On the Capacity and Energy Efficiency of Training-Based Transmissions over Fading Channels

Mustafa Cenk Gursoy
Department of Electrical Engineering
University of Nebraska-Lincoln, Lincoln, NE 68588
Email: gursoy@engr.unl.edu

Abstract

[1] In this paper, the capacity and energy efficiency of training-based communication schemes employed for transmission over a-priori unknown Rayleigh block fading channels are studied. In these schemes, periodically transmitted training symbols are used at the receiver to obtain the minimum mean-square-error (MMSE) estimate of the channel fading coefficients. Initially, the case in which the product of the estimate error and transmitted signal is assumed to be Gaussian noise is considered. In this case, it is shown that bit energy requirements grow without bound as the signal-to-noise ratio (SNR) goes to zero, and the minimum bit energy is achieved at a nonzero SNR value below which one should not operate. The effect of the block length on both the minimum bit energy and the SNR value at which the minimum is achieved is investigated. Energy efficiency analysis is also carried out when peak power constraints are imposed on pilot signals. Flash training and transmission schemes are analyzed and shown to improve the energy efficiency in the low-SNR regime.

In the second part of the paper, the capacity and energy efficiency of training-based schemes are investigated when the channel input is subject to peak power constraints. The capacity-achieving input structure is characterized and the magnitude distribution of the optimal input is shown to be discrete with a finite number of mass points. The capacity, bit energy requirements, and optimal resource allocation strategies are obtained through numerical analysis. The bit energy is again shown to grow without bound as SNR decreases to zero due to the presence of peakedness constraints. Capacity and energy-per-bit are also analyzed under the assumptions that the transmitter interleaves the data symbols before transmission over the channel, and per-symbol peak power constraints are imposed. The improvements in energy efficiency when on-off keying with fixed peak power and vanishing duty cycle is employed are studied. Comparisons of the performances of training-based and noncoherent transmission schemes are provided.

Index Terms: Channel capacity, energy-per-bit, energy efficiency, training-based transmission, capacity-achieving input distribution, optimal resource allocation, Rayleigh block fading channels, channel estimation.

[1]This work was supported in part by the NSF CAREER Grant CCF-0546384. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Nice, France, June 2007.

I. Introduction

In wireless communications, channel conditions vary randomly over time due to mobility and changing environment, and the degree of channel side information (CSI) assumed to be available at the receiver and transmitter is a key assumption in the study of wireless fading channels. The case in which the channel is assumed to be perfectly known at the receiver and/or transmitter has been extensively studied. In an early work, Ericsson [1] obtained the capacity of flat fading channels with perfect receiver CSI. More recently, Ozarow *et al.* [2] studied the average and outage capacity values in the cellular mobile radio setting assuming perfect channel knowledge at the receiver. Goldsmith and Varaiya [3] analyzed the capacity of flat fading channels with perfect CSI at the transmitter and/or receiver.

The assumption of having perfect channel knowledge is unwarranted when communication is trying to be established in a highly mobile environment. This consideration has led to another line of work where both the receiver and transmitter are assumed to be completely uninformed of the channel conditions. Abou-Faycal *et al.* [4] studied the capacity of the unknown Rayleigh fading channel and showed that the optimal input amplitude has a discrete structure. This is in stark contrast to the optimality of a continuous Gaussian input in known channels. In [16] and [18], the discreteness of the capacity-achieving amplitude distribution is proven for noncoherent Rician fading channels under input peakedness constraints. When the input is subject to peak power constraints, the discrete nature of the optimal input is shown for a general class of single-input single-output channels in [7]. Marzetta and Hochwald [5] gave a characterization of the optimal input structure for unknown multiple-antenna Rayleigh fading channels. This analysis subsequently led to the proposal of unitary space-time modulation techniques [6]. Chan *et al* [8] considered conditionally Gaussian multiple-input multiple-output (MIMO) channels with bounded inputs and proved the discreteness of the optimal input under certain conditions. Zheng and Tse [10] analyzed the multiple-antenna Rayleigh channels and identified the high signal-to-noise ratio (SNR) behavior of the channel capacity.

Heretofore, the two extreme assumptions of having either perfect CSI or no CSI have been discussed. Practical wireless systems live in between these two extremes. Unless there is very high mobility, wireless systems generally employ estimation techniques to learn the channel conditions, albeit with errors. Hence, it is of utmost interest to analyze fading channels with imperfect CSI. Médard [13] investigated the effect upon channel capacity of imperfect channel knowledge and obtained upper and lower bounds on the input-output mutual information. Lapidoth and Shamai [12] analyzed the effects of channel estimation errors on the performance if Gaussian codebooks are used and nearest neighbor decoding is employed. The capacity of imperfectly-known fading channels is characterized in the low-SNR regime in [14] and in the high-SNR regime in [9].

The aforementioned studies have not considered explicit training and estimation techniques, and resources allocated to them. Recently, Hassibi and Hochwald [23] studied training schemes to learn the multiple-antenna

channels. In this work, power and time dedicated to training is optimized by maximizing a lower bound on the capacity. Similar training techniques are also discussed in [10]. Due to its practical significance, the information-theoretic analysis of training schemes has attracted much interest (see e.g., [24]-[35]). Since exact capacity expressions are difficult to find, these studies have optimized the training signal power, duration, and placement using capacity bounds. Since Gaussian noise is the worst-case uncorrelated additive noise in a Gaussian setting [23], a capacity lower bound is generally obtained by assuming the product of the estimate error and the transmitted signal as another source of Gaussian noise. In the above cited work, training symbols are employed to solely facilitate channel estimation. However, we note that training symbols can also be used for timing- and frequency-offset synchronization, and channel equalization [36]-[38]. Tong *et al.* in [22] present an overview of pilot-assisted wireless transmissions and discuss design issues from both information-theoretic and signal processing perspectives.

Another important concern in wireless communications is the efficient use of limited energy resources. In systems where energy is at a premium, minimizing the energy cost per unit transmitted information will improve the efficiency. Hence, the energy required to reliably send one bit is a metric that can be adopted to measure the performance. Generally, energy-per-bit requirement is minimized, and hence the energy efficiency is maximized, if the system operates in the low-SNR regime. In [14], Verdú has analyzed the tradeoff between the spectral efficiency and bit energy in the low-SNR regime for a general class of channels and shown that the normalized received minimum bit energy of -1.59 dB is achieved as $\text{SNR} \to 0$ in averaged power limits channels regardless of the availability of CSI at the receiver. On the other hand, [14] has proven that if the receiver has imperfect CSI, the wideband slope, which is the slope of the spectral efficiency curve at zero spectral efficiency, is zero. Hence, approaching the minimum bit energy of -1.59 dB is extremely slow, and moreover it requires input signals with increasingly higher peak-to-average power ratios. The impact upon the energy efficiency of limiting the peakedness of signals is analyzed in [17]. The wideband channel capacity in the presence of input peakedness constraints is investigated in [15], [19], and [20].

Energy efficiency, which is of paramount importance in many wireless systems, has not been the core focus of the aforementioned work on training schemes. Moreover, previous studies optimized the training parameters by using capacity lower bounds. These achievable rate expressions are relevant for systems in which the channel estimate is assumed to be perfect and transmission and reception is designed for a known channel. Note that these assumptions will lead to poor performance unless the SNR is high or the channel coherence time is long.

The contributions of this paper are the following:

- We provide an energy efficiency perspective by analyzing the performance of training techniques in the low-SNR regime. Note that at low SNR levels, the quality of the channel estimate is far from being perfect. We quantify the performance losses in terms of energy efficiency in the worst-case scenario where the estimate

is assumed to be perfect. We identify an SNR level below which one should avoid operating. We consider flash training and transmission techniques to improve the performance.

- We obtain the exact capacity of training-based schemes by characterizing the structure of the capacity-achieving input distribution under input peak power constraints which are highly relevant in practical applications. Optimal resource allocation is performed using the exact capacity values. Improvements in energy efficiency with respect to the worst-case scenario are shown.
- We compare the performances of untrained noncoherent and training-based communication schemes under peak power constraints and show through numerical results that performance loss experienced by training-based schemes is small even at low SNR levels and small values of coherence time. On the other hand, if data symbols are interleaved and experience independent fading, we show that training-based schemes outperform noncoherent techniques.
- We find the attainable bit energy levels in the low-SNR regime when limitations on the peak-to-average power ratio are relaxed and on-off keying with fixed power and vanishing duty cycle is used to transmit information.

The organization of the paper is as follows. Section II provides the channel model. In Section III, training-based transmission and reception is described. In Section IV, we study the achievable rates and energy efficiency in the case where the product of the channel estimate and the transmitted signal is assumed to be Gaussian noise. In Section V, we analyze the capacity and the energy efficiency of training-based schemes when the input is subject to peak power limitations. Section VI includes our conclusions. Proofs of several results are relegated to the Appendix.

II. Channel Model

We consider Rayleigh block-fading channels where the input-output relationship within a block of m symbols is given by

$$\mathbf{y} = h\mathbf{x} + \mathbf{n} \tag{1}$$

where $h \sim \mathcal{CN}(0, \gamma^2)$ [2] is a zero-mean circularly symmetric complex Gaussian random variable with variance $E\{|h|^2\} = \gamma^2$, and $\mathbf{n}$ is a zero-mean, m complex-dimensional Gaussian random vector[3] with covariance matrix $E\{\mathbf{n}\mathbf{n}^\dagger\} = N_0\mathbf{I}$. $\mathbf{x}$ and $\mathbf{y}$ are the m complex-dimensional channel input and output vectors, respectively. It is assumed that the fading coefficients stay constant for a block of m symbols and have independent realizations

[2] $\mathbf{x} \sim \mathcal{CN}(\mathbf{d}, \mathbf{\Sigma})$ is used to denote that $\mathbf{x}$ is a complex Gaussian random vector with mean $E\{\mathbf{x}\} = \mathbf{d}$ and covariance $E\{(\mathbf{x}-\mathbf{d})(\mathbf{x}-\mathbf{d})^\dagger\} = \mathbf{\Sigma}$

[3] Note that in the channel model (1), $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{n}$ are column vectors.

for each block. It is further assumed that neither the transmitter nor the receiver has prior knowledge of the realizations of the fading coefficients.

III. Training-Based Transmission and Reception

We assume that pilot symbols are employed in the system to facilitate channel estimation at the receiver. Hence, the system operates in two phases, namely training and data transmission. In the training phase, pilot symbols known at the receiver are sent from the transmitter and the received signal is

$$\mathbf{y}_t = h\mathbf{x}_t + \mathbf{n}_t \tag{2}$$

where $\mathbf{y}_t$, $\mathbf{x}_t$, and $\mathbf{n}_t$ are l-dimensional vectors signifying the fact that l out of m input symbols are devoted to training. It is assumed that the receiver employs minimum mean-square error (MMSE) estimation to obtain the estimate

$$\hat{h} = E\{h|\mathbf{y}_t\} = \frac{\gamma^2}{\gamma^2\|\mathbf{x}_t\|^2 + N_0}\mathbf{x}_t^\dagger \mathbf{y}_t. \tag{3}$$

With this estimate, the fading coefficient can now be expressed as

$$h = \hat{h} + \tilde{h} \tag{4}$$

where

$$\hat{h} \sim \mathcal{CN}\left(0, \frac{\gamma^4\|\mathbf{x}_t\|^2}{\gamma^2\|\mathbf{x}_t\|^2 + N_0}\right) \quad \text{and} \quad \tilde{h} \sim \mathcal{CN}\left(0, \frac{\gamma^2 N_0}{\gamma^2\|\mathbf{x}_t\|^2 + N_0}\right). \tag{5}$$

Note that $\tilde{h}$ denotes the error in the channel estimate. Following the training phase, the transmitter sends the $(m-l)$-dimensional data vector $\mathbf{x}_d$, and the receiver equipped with the knowledge of the channel estimate operates on the received signal

$$\mathbf{y}_d = \hat{h}\mathbf{x}_d + \tilde{h}\mathbf{x}_d + \mathbf{n}_d \tag{6}$$

to recover the transmitted information. We note that since training-based schemes are studied in this paper, memoryless fading channels in which $m = 1$ are not considered, and it is assumed throughout the paper that the block length satisfies $m \geq 2$.

IV. Achievable Rates and Energy Efficiency in the Worst Case Scenario

A. Average Power Limited Case

In this section, we assume that the input is subject to an average power constraint

$$E\{\|\mathbf{x}\|^2\} \leq mP. \tag{7}$$

Our overall goal is to identify the bit energy values that can be attained with optimized training parameters such as the power and duration of pilot symbols. The least amount of energy required to send one information bit reliably is given by[4]

$$\frac{E_b}{N_0} = \frac{\text{SNR}}{C(\text{SNR})} \tag{8}$$

where $C(\text{SNR})$ is the channel capacity in bits/symbol. In this section, we follow the general approach in the literature and consider a lower bound on the channel capacity by assuming that

$$\mathbf{z} = \tilde{h}\mathbf{x}_d + \mathbf{n}_d \tag{9}$$

is a Gaussian noise vector that has a covariance of

$$E\{\mathbf{z}\mathbf{z}^\dagger\} = \sigma_{\tilde{h}}^2 E\{\mathbf{x}_d\mathbf{x}_d^\dagger\} + N_0\mathbf{I}, \tag{10}$$

and is uncorrelated with the input signal $\mathbf{x}_d$. With this assumption, the channel model becomes

$$\mathbf{y}_d = \hat{h}\mathbf{x}_d + \mathbf{z}. \tag{11}$$

This model is called the worst-case scenario since the channel estimate is assumed to be perfect, and the noise is modeled as Gaussian, which presents the worst case [23]. The capacity of the channel in (11), which acts as a lower bound on the capacity of the channel in (6), is achieved by a Gaussian input with

$$E\{\mathbf{x}_d\mathbf{x}_d^\dagger\} = \frac{(1-\delta^*)mP}{m-1}\mathbf{I} \tag{12}$$

where δ^* is the optimal fraction of power allocated to the pilot symbol, i.e., $|x_t|^2 = \delta^* mP$. The optimal value is given by

$$\delta^* = \sqrt{\eta(\eta+1)} - \eta \tag{13}$$

where

$$\eta = \frac{m\,\text{SNR} + (m-1)}{m(m-2)\text{SNR}} \quad \text{and} \quad \text{SNR} = \frac{\gamma^2 P}{N_0}. \tag{14}$$

Note that SNR in (14) is the received signal-to-noise ratio. In the average power limited case, sending a single pilot is optimal because instead of increasing the number of pilot symbols, a single pilot with higher power can be used and a decrease in the duration of the data transmission can be avoided. Hence, the optimal $\mathbf{x}_d$ is an

[4]Note that $\frac{E_b}{N_0}$ is the bit energy normalized by the noise power spectral level N_0.

$(m-1)$-dimensional Gaussian vector. Since the above results are indeed special cases of those in [23], the details are omitted. The resulting capacity expression[5] is

$$\begin{aligned} C_L(\text{SNR}) &= \frac{m-1}{m} E_w \left\{ \log \left(1 + \frac{\phi(\text{SNR})\text{SNR}^2}{\psi(\text{SNR})\text{SNR} + (m-1)} |w|^2 \right) \right\} \\ &= \frac{m-1}{m} E_w \left\{ \log \left(1 + f(\text{SNR})|w|^2 \right) \right\} \text{ nats/symbol} \end{aligned} \tag{15}$$

where

$$\phi(\text{SNR}) = \delta^*(1-\delta^*)m^2, \quad \text{and} \quad \psi(\text{SNR}) = (1+(m-2)\delta^*)m, \tag{16}$$

and $w \sim \mathcal{CN}(0,1)$. Note also that the expectation in (15) is with respect to the random variable w. The bit energy values in this setting are given by

$$\frac{E_{b,U}}{N_0} = \frac{\text{SNR}}{C_L(\text{SNR})} \log 2 \tag{17}$$

where C_L is in nats/symbol. $\frac{E_{b,U}}{N_0}$ provides the least amount of normalized bit energy values in the worst-case scenario and also serves as an upper bound on the achievable bit energy levels of channel (6). It is shown in [12] that if the channel estimate is assumed to be perfect, and Gaussian codebooks designed for known channels are used, and scaled nearest neighbor decoding is employed at the receiver, then the generalized mutual information has an expression similar to (15) (see [12, Corollary 3.0.1]). Hence $\frac{E_{b,U}}{N_0}$ also gives a good indication of the energy requirements of a system operating in this fashion. The next result provides the asymptotic behavior of the bit energy as SNR decreases to zero.

Proposition 1: The normalized bit energy (17) grows without bound as the signal-to-noise ratio decreases to zero, i.e.,

$$\left.\frac{E_{b,U}}{N_0}\right|_{C_L=0} = \lim_{\text{SNR}\to 0} \frac{\text{SNR}}{C_L(\text{SNR})} \log 2 = \frac{\log 2}{\dot{C}_L(0)} = \infty. \tag{18}$$

Proof: In the low SNR regime, we have

$$C_L(\text{SNR}) = \frac{m-1}{m} \left(f(\text{SNR}) E\{|w|^2\} + o(f(\text{SNR})) \right) \tag{19}$$

$$= \frac{m-1}{m} \left(f(\text{SNR}) + o(f(\text{SNR})) \right). \tag{20}$$

As SNR $\to 0$, $\delta^* \to 1/2$, and hence $\phi(\text{SNR}) \to m^2/4$ and $\psi(\text{SNR}) \to m + m(m-2)/2$. Therefore, it can easily be seen that

$$f(\text{SNR}) = \frac{m^2}{4(m-1)} \text{SNR}^2 + o(\text{SNR}^2) \tag{21}$$

from which we have $\dot{C}_L(0) = 0$. □

[5]Unless specified otherwise, all logarithms are to the base e.

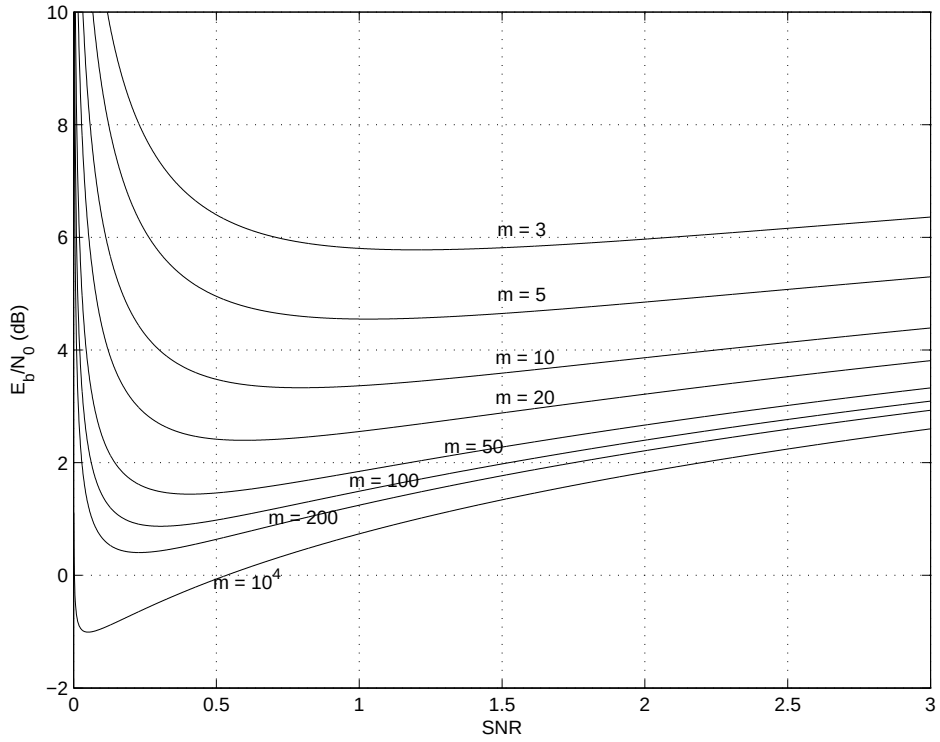

Fig. 1. Energy per bit $E_{b,U}/N_0$ vs. SNR in the worst-case scenario

The fact that C_L decreases as SNR^2 as SNR goes to zero has already been pointed out in [23]. The reason for this behavior is that as SNR decreases, the power of $\hat{h}$ (5) decreases linearly with SNR and hence the quality of the channel estimate deteriorates. Since the channel estimate is assumed to be perfect, the effective signal-to-noise ratio decays as SNR^2 leading to the observed result. Proposition 1 shows the impact of this behavior on the energy-per-bit, and indicates that it is extremely energy-inefficient to operate at very low SNR values. The result holds regardless of the size of the block length m as long as it is finite. We further conclude that in a training-based scheme where the channel estimate is assumed to be perfect, the minimum energy per bit is achieved at a nonzero SNR value. This most energy-efficient operating point can be obtained by numerical analysis. We can easily compute $C_L(\text{SNR})$ in (15), and hence the bit energy values.

Figure 1 plots the normalized bit energy curves as a function of SNR for block lengths of $m = 3, 5, 10, 20, 50, 100, 200, 10^4$. As predicted, for each block length value, the minimum bit energy is achieved at nonzero SNR, and the bit energy requirement increases as $\text{SNR} \to 0$. It is been noted in [23] that training-based schemes, which assume the channel estimate to be perfect, perform poorly at very low SNR values, and the exact transition point below which one should not operate in this fashion is deemed as not clear. Here, we propose the SNR level at which the minimum bit energy is achieved as a transition point since operating below this point results in higher bit energy

requirements. It is further seen in Fig. 1 that the minimum bit energy is attained at an SNR value that satisfies

$$\frac{d}{d\text{SNR}}\left(\frac{E_{b,U}}{N_0}\right) = \frac{d}{d\text{SNR}}\left(\frac{\text{SNR}\log 2}{C_L(\text{SNR})}\right) = 0. \tag{22}$$

Another observation from Fig. 1 is that the minimum bit energy decreases with increasing m and is achieved at a lower SNR value. The following result sheds a light on the asymptotic behavior of the capacity as $m \to \infty$.

Theorem 1: As the block length m increases, C_L approaches to the capacity of the perfectly known channel, i.e.,

$$\lim_{m\to\infty} C_L(\text{SNR}) = E_w\{\log(1 + \text{SNR}|w|^2)\}. \tag{23}$$

Moreover, define $\chi = 1/m$. Then

$$\left.\frac{dC_L(\text{SNR})}{d\chi}\right|_{\chi=0} = -\infty. \tag{24}$$

Proof: We have

$$\begin{aligned}
\lim_{m\to\infty} C_L(\text{SNR}) &= \lim_{m\to\infty} E_w\left\{\log\left(1 + f(\text{SNR})|w|^2\right)\right\} && (25)\\
&= E_w\left\{\lim_{m\to\infty} \log\left(1 + f(\text{SNR})|w|^2\right)\right\} && (26)\\
&= E_w\left\{\log\left(1 + |w|^2 \lim_{m\to\infty} f(\text{SNR})\right)\right\} && (27)\\
&= E_w\left\{\log\left(1 + \text{SNR}|w|^2\right)\right\}. && (28)
\end{aligned}$$

(25) follows from the fact that $(m-1)/m \to 1$ as $m \to \infty$. For (26) to hold, we invoke the Dominated Convergence Theorem [40]. Note that

$$\begin{aligned}
\left|\log(1+f(\text{SNR})|w|^2)\right| &\le f(\text{SNR})|w|^2 && (29)\\
&= \frac{\phi(\text{SNR})\text{SNR}^2}{\psi(\text{SNR})\text{SNR} + (m-1)}|w|^2 && (30)\\
&\le \frac{\phi(\text{SNR})}{\psi(\text{SNR})}\text{SNR}|w|^2 && (31)\\
&= \frac{\delta^*(1-\delta^*)m^2}{m + m(m-2)\delta^*}\text{SNR}|w|^2 && (32)\\
&= \frac{(1-\delta^*)m^2}{\frac{m}{\delta^*} + m(m-2)}\text{SNR}|w|^2 && (33)\\
&\le \frac{(1-\delta^*)m}{m-2}\text{SNR}|w|^2 && (34)\\
&\le 3\text{SNR}|w|^2 \quad \text{for } m \ge 3 && (35)
\end{aligned}$$

where (34) is obtained by removing $\frac{m}{\delta^*}$ in the denominator and (35) follows from the facts that $1-\delta^* \le 1$ and $\frac{m}{m-2} \le 3$ for all $m \ge 3$. If $m = 2$, we have $\phi(\text{SNR}) = 1$, $\psi(\text{SNR}) = 2$, and hence

$$\left|\log(1+f(\text{SNR})|w|^2)\right| = \log\left(1 + \frac{\text{SNR}^2}{2\text{SNR}+1}|w|^2\right) \le \frac{1}{2}\text{SNR}|w|^2. \tag{36}$$

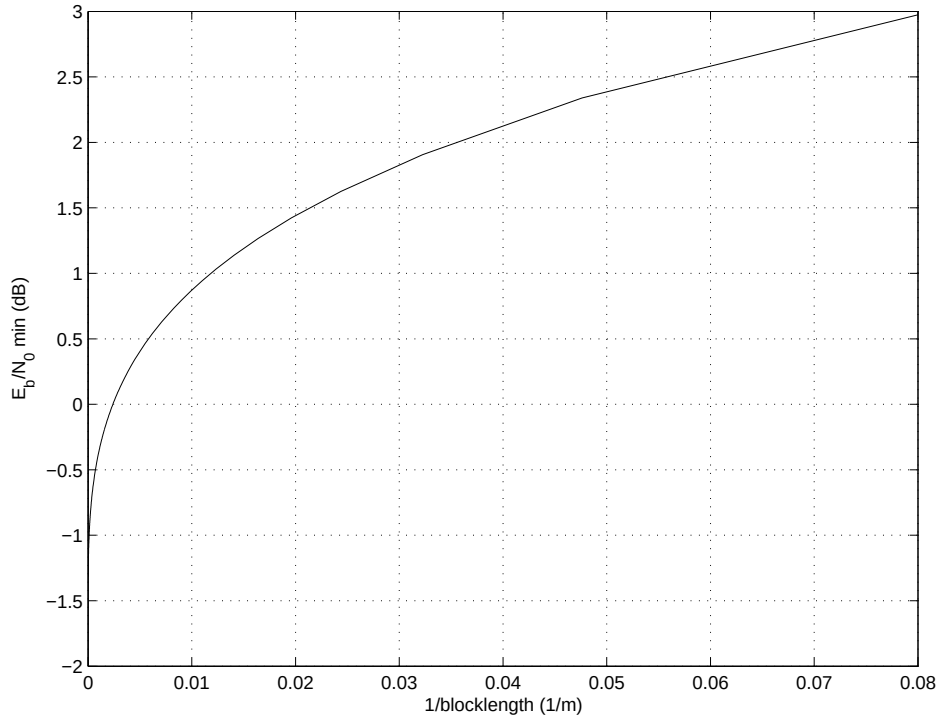

Fig. 2. Minimum energy per bit $\frac{E_{b,U}}{N_0}_{min}$ vs. $\frac{1}{m}$ in the worst-case scenario

Therefore, $3\text{SNR}|w|^2$ is an upper bound that applies for all integer values $m \geq 2$. Furthermore, the upper bound does not depend on m and is integrable, i.e., $E_w\{3\text{SNR}|w|^2\} = 3\text{SNR} < \infty$. Hence, the Dominated Convergence Theorem applies and (26) is justified. (27) is due to the fact that logarithm is a continuous function. (28) can easily be verified by noting that $m^2\delta^*$ is the fastest growing component, increasing as $m^{\frac{3}{2}}$ with increasing m.

(24) follows again from the application of the Dominated Convergence Theorem and the fact that the derivative of $f(\text{SNR})$ with respect to $\chi = 1/m$ at $\chi = 0$ is $-\infty$. □

The first part of Theorem 1 is not surprising and is expected because reference [5] has already shown that as the block length grows, the perfect knowledge capacity is achieved even if no channel estimation is performed. This result agrees with our observation in Fig. 1 that -1.59 dB is approached at lower SNR values as m increases. However, the rate of approach is very slow in terms of the block size, as proven in the second part of Theorem 1 and evidenced in Fig. 2. Due to the infinite slope[6] observed in the figure, approaching -1.59 dB is very demanding in block length.

[6]Note that Theorem 1 implies that the slope of $\frac{\text{SNR}}{C_L(\text{SNR})}$ at $\chi = \frac{1}{m} = 0$ is ∞.

B. Peak Power Constraint on the Pilot

Heretofore, we have assumed that there are no peak power constraints imposed on either the data or pilot symbols. Recall that the power of the pilot symbol is given by

$$|x_t|^2 = \delta^* mP = \sqrt{\xi(\xi + mP)} - \xi \tag{37}$$

where $\xi = \frac{m\gamma^2 P + (m-1)N_0}{(m-2)\gamma^2}$. We immediately observe from (37) that the pilot power increases at least as $\sqrt{m}$ as m increases. For large block sizes, such an increase in the pilot power may be prohibitive in practical systems. Therefore, it is of interest to impose a peak power constraint on the pilot in the following form:

$$|x_t|^2 \leq \kappa P. \tag{38}$$

Since the average power is uniformly distributed over the data symbols, the average power of a data symbol is proportional to P and is at most $(1-\delta^*)2P$ for any block size. Therefore, κ can be seen as a limitation on the peak-to-average power ratio. Note that we will allow Gaussian signaling for data transmission. Hence, there are no hard peak power limitations on data signals. This approach will enable us to work with a closed-form capacity expression. Although Gaussian signals can theoretically assume large values, the probability of such values is decreasing exponentially. The case in which a peak power constraint is imposed on both the training and data symbols is treated in the Section V.

If the optimal power allocated to a single pilot exceeds κP, i.e., $\delta^* mP > \kappa P \Rightarrow \delta^* m > \kappa$, the peak power constraint on the pilot becomes active. In this case, more than just a single pilot may be needed for optimal performance.

In this section, we address the optimization of the number of pilot symbols when each pilot symbol has fixed power $|x_{t,i}|^2 = \kappa P \ \forall i$. If the number of pilot symbols is $l < m$, then $\|\mathbf{x}_t\|^2 = l\kappa P$ and, as we know from Section III,

$$\hat{h} \sim \mathcal{CN}\left(0, \frac{\gamma^4 l\kappa P}{\gamma^2 l\kappa P + N_0}\right) \text{ and } \tilde{h} \sim \mathcal{CN}\left(0, \frac{\gamma^2 N_0}{\gamma^2 l\kappa P + N_0}\right).$$

Similarly as before, when the estimate error is assumed to be another source of additive noise and overall additive noise is assumed to be Gaussian, the input-output mutual information achieved by Gaussian signaling is given by

$$I_{L,p} = \frac{m-l}{m} E_w \left\{\log\left(1 + g(\text{SNR}, l)|w|^2\right)\right\} \tag{39}$$

where $w \sim \mathcal{CN}(0,1)$ and

$$g(\text{SNR}, l) = \frac{l\kappa(m - l\kappa)\text{SNR}^2}{(m - l\kappa + (m-l)l\kappa)\text{SNR} + m - l}. \tag{40}$$

The optimal value of the training duration l that maximizes $I_{L,p}$ can be obtained through numerical optimization.

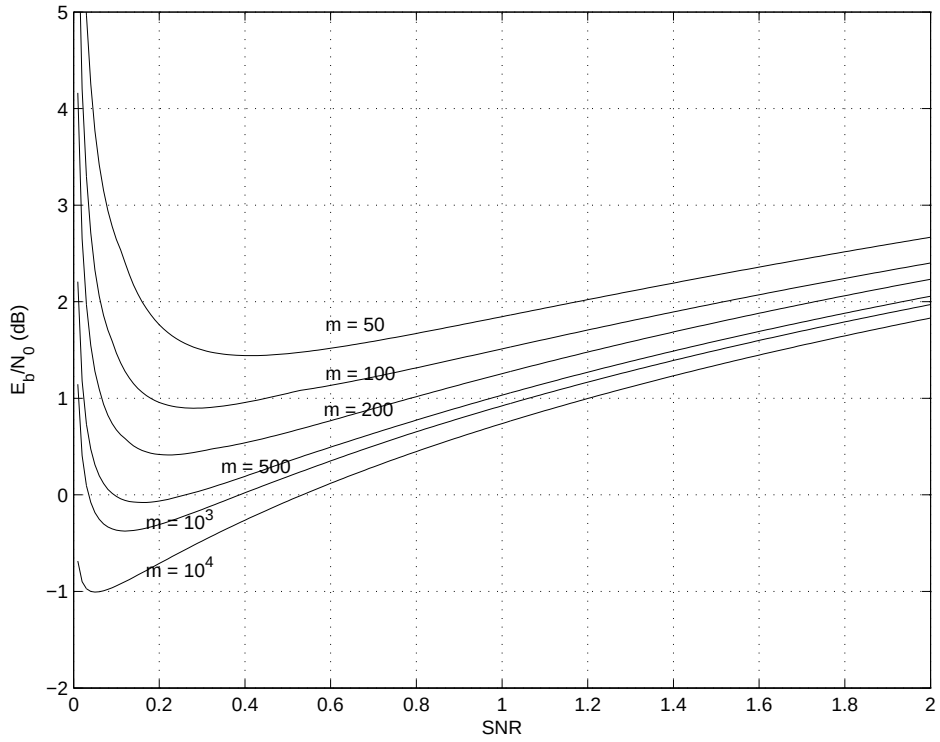

Fig. 3. Energy per bit $E_{b,U}/N_0$ vs. SNR for block sizes of $m = 50, 100, 200, 500, 10^3, 10^4$. The pilot peak power constraint is $|x_t|^2 \leq 10P$.

Fig. 3 plots the normalized bit energy values $\frac{\text{SNR}\log 2}{I_{L,p}}$ in dB obtained with optimal training duration for different block lengths. The peak power constraint imposed on a pilot symbol is $|x_t|^2 \leq 10P$. Fig. 4 gives the optimal number of pilot symbols per block. From Fig. 3, we observe that the minimum bit energy, which is again achieved at a nonzero value of the SNR, decreases with increasing block length and approaches to the fundamental limit of -1.59 dB. We note from Fig. 4 that the number pilot symbols per block increases as the block length increases or as SNR decreases to zero. When there are no peak constraints, $\delta^* \to 1/2$ as SNR $\to 0$. Hence, we need to allocate approximately half of the available total power mP to the single pilot signal in the low-power regime, increasing the peak-to-average power ratio. In the limited peak power case, this requirement is translated to the requirement of more pilot symbols per block at low SNR values.

Table I lists, for different values of m, the minimum bit energy values, the required number of pilot symbols at this level, and the SNR at which minimum bit energy is achieved. It is again assumed that $\kappa = 10$. The last column of the table provides the minimum bit energy attained when there are no peak power constraints on the pilot signal. As the block size increases, the minimum bit energy is achieved at a lower SNR value while a longer training duration is required. Furthermore, comparison with the last column indicates that the loss in minimum

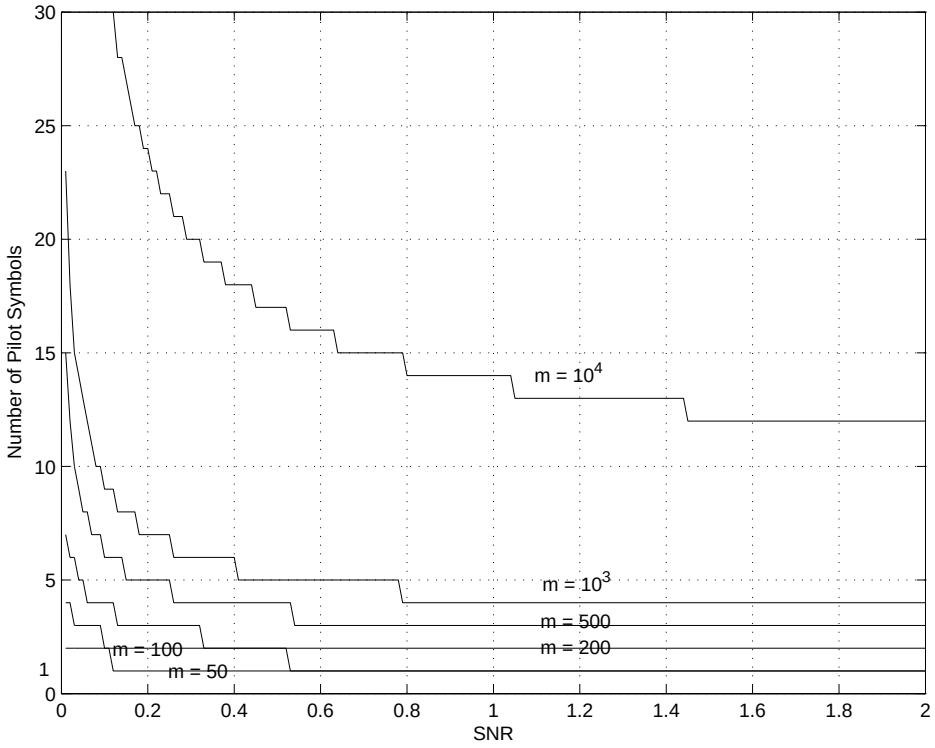

Fig. 4. Number of pilot symbols per block vs. SNR

bit energy incurred by the presence of peak power constraints is negligible. The following result shows that the capacity of the perfectly known channel, and hence the minimum bit energy of -1.59dB, is approached with simultaneous growth of training duration and block length. Note that this result conforms with the results in Table I.

Proposition 2: Assume that the training duration $l(m, \text{SNR})$ increases as m increases and satisfies

$$\lim_{m\to\infty} \frac{l(m, \text{SNR})}{m} = 0. \tag{41}$$

Then, $\lim_{m\to\infty} I_{L,p} = E_w\{\log(1 + \text{SNR}|w|^2)\}$.

Proof: We have

$$\begin{aligned}
\lim_{m\to\infty} I_{L,p} &= \lim_{m\to\infty} \left(1 - \frac{l}{m}\right) E_w\left\{\log\left(1 + g(\text{SNR}, l)|w|^2\right)\right\} \\
&= \lim_{m\to\infty} E_w\left\{\log\left(1 + g(\text{SNR}, l)|w|^2\right)\right\} \qquad (42)\\
&= E_w\left\{\lim_{m\to\infty} \log\left(1 + g(\text{SNR}, l)|w|^2\right)\right\} \qquad (43)\\
&= E_w\{\log(1 + \text{SNR}|w|^2)\}. \qquad (44)
\end{aligned}$$

TABLE I

	$\frac{E_{b,U}}{N_0}_{\min}$(dB)	# of pilots	SNR	$\frac{E_{b,U}}{N_0}_{\min}$ (dB) (no peak constraints)
$m = 50$	1.441	1	0. 41	1.440
$m = 100$	0.897	2	0.28	0.871
$m = 200$	0.413	3	0.22	0.404
$m = 500$	-0.079	5	0.16	- 0.085
$m = 10^3$	-0.375	9	0.12	-0.378
$m = 10^4$	-1.007	44	0.05	-1.008

(42) follows from the condition (41). (43) can be justified by invoking the Dominated Convergence Theorem [40] similarly as in the proof of Theorem 1. (44) follows from

$$\lim_{m\to\infty} g(\text{SNR}, l) = \text{SNR}, \tag{45}$$

which holds if the conditions of the theorem are met. □

C. Flash Training and Transmission

One approach to improve the energy efficiency in the low SNR regime is to increase the peak power of the transmitted signals. This can be achieved by transmitting ν fraction of the time with power P/ν. Note that training also needs to be performed only ν fraction of the time. In this section, no peak power constraints are imposed on pilot symbols. This type of training and communication, called flash training and transmission scheme, is analyzed in [11] where it is shown that the minimum bit energy of -1.59 dB can be achieved if the block length m increases at a certain rate as SNR decreases. In the setting we consider, flash transmission scheme achieves the following rate:

$$C_{fL}(\text{SNR}, \nu) = \nu(\text{SNR}) C_L\left(\frac{\text{SNR}}{\nu(\text{SNR})}\right) \tag{46}$$

where $0 < \nu(\cdot) \leq 1$ is the duty cycle which in general is a function of the SNR. First, we show that flash transmission using peaky Gaussian signals does not improve the minimum bit energy.

Proposition 3: For any duty cycle function $\nu(\cdot)$,

$$\inf_{\text{SNR}} \frac{\text{SNR}}{C_{fL}(\text{SNR}, \nu)} \geq \inf_{\text{SNR}} \frac{\text{SNR}}{C_L(\text{SNR})}. \tag{47}$$

Proof: Note that for any SNR and $\nu(\text{SNR})$,

$$\frac{\text{SNR}}{C_{fL}(\text{SNR},\nu)} = \frac{\frac{\text{SNR}}{\nu(\text{SNR})}}{C_L\left(\frac{\text{SNR}}{\nu(\text{SNR})}\right)} = \frac{\tilde{\text{SNR}}}{C_L(\tilde{\text{SNR}})} \geq \inf_{\text{SNR}} \frac{\text{SNR}}{C_L(\text{SNR})} \tag{48}$$

where $\tilde{\text{SNR}}$ is defined as the new SNR level. Since the inequality in (48) holds for any SNR and $\nu(\cdot)$, it also holds for the infimum of the left-hand side of (48), and hence the result follows. □

We classify the duty cycle function into three categories:

1) $\nu(\cdot)$ that satisfies $\lim_{\text{SNR}\to 0} \frac{\text{SNR}}{\nu(\text{SNR})} = 0$
2) $\nu(\cdot)$ that satisfies $\lim_{\text{SNR}\to 0} \frac{\text{SNR}}{\nu(\text{SNR})} = \infty$
3) $\nu(\cdot)$ that satisfies $\lim_{\text{SNR}\to 0} \frac{\text{SNR}}{\nu(\text{SNR})} = a$ for some constant $a > 0$.

Next, we analyze the performance of each category of duty cycle functions in the low-SNR regime.

Theorem 2: If $\nu(\cdot)$ is chosen from either Category 1 or 2,

$$\left.\frac{E_{b,U}}{N_0}\right|_{C_{fL}=0} = \lim_{\text{SNR}\to 0} \frac{\text{SNR}}{C_{fL}(\text{SNR},\nu)} \log 2 = \infty. \tag{49}$$

If $\nu(\cdot)$ is chosen from Category 3,

$$\left.\frac{E_{b,U}}{N_0}\right|_{C_{fL}=0} = \frac{m}{m-1} \frac{a}{E_w\{\log_2(1+f(a)|w|^2)\}}. \tag{50}$$

Proof: We first note that by Jensen's inequality,

$$\frac{C_{fL}(\text{SNR},\nu)}{\text{SNR}} \leq \frac{m-1}{m} \frac{\nu(\text{SNR})}{\text{SNR}} \log\left(1 + f\left(\frac{\text{SNR}}{\nu(\text{SNR})}\right)\right) \tag{51}$$

$$\stackrel{\text{def}}{=} \zeta(\text{SNR},\nu). \tag{52}$$

First, we consider category 1. In this case, as $\text{SNR} \to 0$, $\frac{\text{SNR}}{\nu(\text{SNR})} \to 0$. As shown before, the logarithm in (51) scales as $\frac{\text{SNR}^2}{\nu(\text{SNR}^2)}$ as $\text{SNR} \to 0$, and hence $\zeta(\text{SNR},\nu)$ scales as $\frac{\text{SNR}}{\nu(\text{SNR})}$ leading to

$$\lim_{\text{SNR}\to 0} \frac{C_{fL}(\text{SNR},\nu)}{\text{SNR}} \leq \lim_{\text{SNR}\to 0} \zeta(\text{SNR},\nu) = 0. \tag{53}$$

In category 2, $\frac{\text{SNR}}{\nu(\text{SNR})}$ grows to infinity as $\text{SNR} \to 0$. Since the $\log(\cdot)$ function on the right hand side of (51) increases only logarithmically as $\frac{\text{SNR}}{\nu(\text{SNR})} \to \infty$, we can easily verify that

$$\lim_{\text{SNR}\to 0} \frac{C_{fL}(\text{SNR},\nu)}{\text{SNR}} \leq \lim_{\text{SNR}\to 0} \zeta(\text{SNR},\nu) = 0. \tag{54}$$

In category 3, $\nu(\text{SNR})$ decreases at the same rate as SNR. In this case, we have

$$\lim_{\text{SNR}\to 0} \frac{C_{fL}(\text{SNR},\nu)}{\text{SNR}} = \lim_{n\to\infty} \frac{C_{fL}\left(\frac{1}{n},\nu\right)}{\frac{1}{n}} \tag{55}$$

$$= \frac{\frac{m-1}{m} E_w\{\lim_{n\to\infty} \log\left(1+f(\frac{1}{n\nu})|w|^2\right)\}}{a} \tag{56}$$

$$= \frac{\frac{m-1}{m} E_w\{\log\left(1+f(a)|w|^2\right)\}}{a} \tag{57}$$

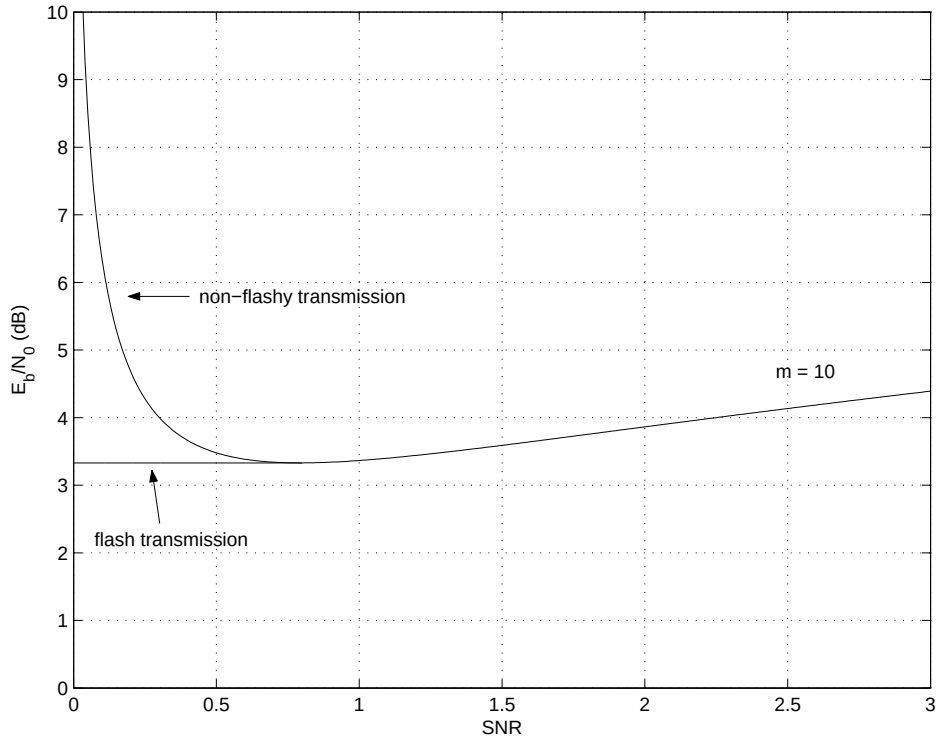

Fig. 5. Energy per bit $E_{b,U}/N_0$ vs. SNR for non-flashy and flash transmissions.

(56) is justified by invoking the Dominated Convergence Theorem and noting the integrable upper bound

$$\left|\log\left(1+f\left(\frac{1}{n\nu}\right)|w|^2\right)\right| \leq 3\frac{1}{n\nu}|w|^2 \leq \frac{3}{\nu}|w|^2 \text{ for } n \geq 1.$$

The above upper bound is given in the proof of Theorem 1. Finally, (57) follows from the continuity of the logarithm. □

Theorem 2 shows that if the rate of the decrease of the duty cycle is faster or slower than SNR as SNR $\to 0$, the bit energy requirement still increases without bound in the low-SNR regime. This observation is tightly linked to the fact that the capacity curve C_L has a zero slope as both SNR $\to 0$ and SNR $\to \infty$. For improved performance in the low-SNR regime, it is required that the duty cycle scale as SNR. A particularly good choice is

$$\nu(\text{SNR}) = \frac{1}{a^*}\text{SNR}$$

where a^* is equal to the SNR level at which the minimum bit energy is achieved in a non-flashy transmission scheme. With this choice, we basically perform time-sharing between SNR $= 0$ and SNR $= a^*$. Fig. 5 plots the normalized bit energy $\frac{E_{b,U}}{N_0}$ as a function of SNR for block size $m = 10$. The minimum bit energy is achieved at SNR $= 0.8$. For SNR < 0.8, flash transmission is employed with $\nu(\text{SNR}) = 1/0.8\,\text{SNR}$. As observed in the figure, the minimum bit energy level can be maintained for lower values of SNR at the cost of increased peak-to-average

power ratio. It should be noted that the optimal point of operation is still at $\text{SNR} = 0.8$ since operating at $\text{SNR} < 0.8$ will result in reduced data rates without any improvements in the bit energy. From a different perspective, if SNR is the signal-to-noise ratio per unit bandwidth, then increasing the bandwidth so that $\text{SNR} < 0.8$ will not produce any energy savings. However, in circumstances in which regulations or device properties dictate operation at SNR values lower than the minimum bit energy point, flash transmission can be adopted to improve the energy efficiency.

V. Capacity and Energy Efficiency in the Presence of Peak Power Limitations

In this section, we consider the channel

$$\mathbf{y}_d = \hat{h}\mathbf{x}_d + \tilde{h}\mathbf{x}_d + \mathbf{n}_d \tag{58}$$

and assume that the channel input is subject to the following peak power constraint

$$\|\mathbf{x}\|^2 \overset{\text{a.s.}}{\leq} mP. \tag{59}$$

In this setting, it is again easy to see that the transmission of a single pilot is optimal. Since the peak power constraint is imposed on the input vector $\mathbf{x}$, the pilot power can be varied instead of increasing the number of pilot symbols. Similarly as before, we assume that the pilot symbol power is

$$|x_t|^2 = \delta mP. \tag{60}$$

Therefore, the $(m-1)$-dimensional data vector $\mathbf{x}_d$ is subject to

$$\|\mathbf{x}_d\|^2 \overset{\text{a.s.}}{\leq} (1-\delta)mP. \tag{61}$$

Our goal is to solve the maximization problem

$$C = \sup_{\delta\in(0,1)} \sup_{\substack{\mathbf{x}_d \\ \|\mathbf{x}_d\|^2 \overset{\text{a.s.}}{\leq} (1-\delta)mP}} \frac{1}{m} I(\mathbf{x}_d; \mathbf{y}_d|\hat{h}) \tag{62}$$

and obtain the channel capacity, and identify the capacity-achieving input distribution and the optimal value of the power allocation coefficient δ. The input-output mutual information is

$$I(\mathbf{x}_d; \mathbf{y}_d|\hat{h}) = E_{\hat{h}} E_{\mathbf{x}_d} \int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h}) \log \frac{f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})}{f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})}\, d\mathbf{y} \tag{63}$$

where

$$f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h}) = \frac{\exp\left(-(\mathbf{y}-\hat{h}\mathbf{x}_d)^\dagger(\tilde{\gamma}^2\mathbf{x}_d\mathbf{x}_d^\dagger + N_0\mathbf{I})^{-1}(\mathbf{y}-\hat{h}\mathbf{x}_d)\right)}{\pi^{m-1}N_0^{m-2}(\tilde{\gamma}^2\|\mathbf{x}_d\|^2 + N_0)} \tag{64}$$

and

$$\tilde{\gamma}^2 = E\{|\tilde{h}|^2\} = \frac{\gamma^2 N_0}{\gamma^2\delta mP + N_0}. \tag{65}$$

First, we have the following preliminary result on the structure of the capacity-achieving input distribution.

Theorem 3: For the block fading channel (58) where the input is subject to a peak power limitation (61), the capacity-achieving input vector can be written as $\mathbf{x}_d = \|\mathbf{x}_d\|\mathbf{v}$ where $\|\mathbf{x}_d\|$ is a nonnegative real random variable and $\mathbf{v}$ is an independent isotropically distributed unit random vector.

Proof: The proof follows primarily from the same techniques developed in [5]. First note the invariance of the peak constraint (61) to rotations of the input. Since $f_{\mathbf{y}|\mathbf{x}}(\Phi\mathbf{y}|\Phi\mathbf{x}_d, \hat{h}) = f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d, \hat{h})$ for any $(m-1)\times(m-1)$ dimensional deterministic unitary matrix Φ, it can be easily seen that the mutual information is also invariant to deterministic rotations of the input, and the result follows from the concavity of the mutual information which implies that there is no loss in optimality if one uses circularly symmetric input distributions. □

With this characterization, the problem has been reduced to the optimization of the input magnitude distribution, $F_{\mathbf{x}_d}$. We first obtain an equivalent expression for the mutual information when the the input vector has the structure described in Theorem 3.

Theorem 4: When the input is $\mathbf{x}_d = \|\mathbf{x}_d\|\mathbf{v}$ where $\mathbf{v}$ is an isotropically distributed unit vector that is independent of the magnitude $\|\mathbf{x}_d\|$, the input-output mutual information of the channel (58) can be expressed as

$$I(\mathbf{x}_d; \mathbf{y}_d|\hat{h}) = I(F_r|\hat{h}) = -E_{\mathsf{K},r}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) \log\, g(R, F_r, \mathsf{K})\, dR\right\} - E_r\{\log(1+r^2)\} - (m-1) \quad (66)$$

where

$$f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) = \begin{cases} \frac{R^{m-2}}{(m-3)!}\frac{e^{-R-\frac{\mathsf{K}r^2}{1+r^2}}}{1+r^2}\int_0^1 (1-a)^{m-3} e^{\frac{ar^2R}{1+r^2}} I_0\left(\frac{2\sqrt{\mathsf{K}R}r\sqrt{a}}{1+r^2}\right) da & m \geq 3 \\ \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{1+r^2} I_0\left(\frac{2\sqrt{\mathsf{K}R}r}{1+r^2}\right) & m = 2 \end{cases} \quad (67)$$

and

$$g(R, F_r, \mathsf{K}) = \frac{(m-2)!}{R^{m-2}}\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K})\, dF_r. \quad (68)$$

In the above formulations, $R = \frac{\|\mathbf{y}\|^2}{N_0}$, $r = \frac{\tilde{\gamma}\|\mathbf{x}_d\|}{\sqrt{N_0}}$, and $\mathsf{K} = \frac{|\hat{h}|^2}{\tilde{\gamma}^2}$. Furthermore, F_r denotes the distribution function of r. K is an exponential random variable with mean $E\{\mathsf{K}\} = \frac{E\{|\hat{h}^2|\}}{\tilde{\gamma}^2} = \frac{\gamma^2\delta mP}{N_0}$. $E_{\mathsf{K},r}$ denotes the expectation with respect to K and r.

Proof: See Appendix A.

Note that the integral in the mutual information expression in (63) is in general an $2(m-1)$-fold integral. In (66), this has been reduced to a double integral providing a significant simplification especially for numerical analysis. With this result, the channel capacity in nats per symbol can now be reformulated as

$$C = \sup_{\delta\in(0,1)} C_\delta = \sup_{\delta\in(0,1)} \sup_{\substack{F_r \\ r \overset{\text{a.s.}}{\leq} \sqrt{L}}} \frac{1}{m} I(F_r|\hat{h}) \quad (69)$$

where $L = \frac{\gamma^2(1-\delta)mP}{\gamma^2\delta mP+N_0}$. Hence, the capacity is obtained through the optimal choices of the power allocation coefficient δ and normalized input magnitude distribution F_r. Since the inner maximization is over a continuous alphabet, the existence of the capacity-achieving distribution F_r is not guaranteed. Next, we prove the existence of a capacity-achieving input distribution and provide a sufficient and necessary condition for an input to be optimal.

Theorem 5: Fix the value of $\delta \in (0,1)$ and consider the inner maximization in (69). There exists an input distribution F_r that maximizes the mutual information $I(F_r|\hat{h})$. Moreover, an input distribution F_r is capacity-achieving if and only if the following Kuhn-Tucker condition is satisfied:

$$\Phi(r) = E_{\mathsf{K}}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) \log\, g(R,F_r,\mathsf{K})\, dR\right\} + \log(1+r^2) + mC_\delta + (m-1) \geq 0 \quad \forall r \in [0,\sqrt{L}] \tag{70}$$

with equality at the points of increase of F_r[7]. In the above condition, C_δ denotes the result of the inner maximization in (69) .

Proof: See Appendix B.

Having shown the existence of the capacity-achieving input distribution and a sufficient and necessary condition for an input distribution to be optimal, we turn our attention to the characterization of the optimal input.

Theorem 6: Fix the value of $\delta \in (0,1)$. The input distribution that maximizes the mutual information $I(F_r|\hat{h})$ is discrete with a finite number of mass points

Proof: The following upper bound is obtained in Appendix B:

$$g(R,F,\mathsf{K}) \leq (m-2)\, e^{-\frac{R}{1+L}+\sqrt{\mathsf{K}R}} \tag{71}$$

Using this upper bound, we have

$$E_{\mathsf{K}}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) \log\, g(R,F_r,\mathsf{K})\, dR\right\} = E_{\mathsf{K}}E_{R|r,\mathsf{K}}\{\log\, g(R,F_r,\mathsf{K})\} \tag{72}$$

$$\leq \log(m-2) - E_{\mathsf{K}}E_{R|r,\mathsf{K}}\left\{\frac{R}{1+L}\right\} + E_{\mathsf{K}}E_{R|r,\mathsf{K}}\left\{\sqrt{\mathsf{K}R}\right\} \tag{73}$$

$$\leq \log(m-2) - E_{\mathsf{K}}E_{R|r,\mathsf{K}}\left\{\frac{R}{1+L}\right\} + E_{\mathsf{K}}\left\{\sqrt{\mathsf{K}}\sqrt{E_{R|r,\mathsf{K}}\{R\}}\right\} \tag{74}$$

$$\leq \log(m-2) - \frac{(1+E_{\mathsf{K}}\{\mathsf{K}\})r^2+m-1}{1+L} + E_{\mathsf{K}}\left\{\sqrt{\mathsf{K}}\sqrt{(1+\mathsf{K})r^2+m-1}\right\}. \tag{75}$$

(73) follows from (71), and (74) follows from the fact that $E\{\sqrt{R}\} \leq \sqrt{E\{R\}}$. Finally, (75) is obtained by noting that $E_{R|r,\mathsf{K}}\{R\} = (1+\mathsf{K})r^2 + m - 1$. Note that the upper bound in (75), and hence the left-hand-side of (70), decreases to $-\infty$ as $r \to \infty$ due to the presence of $-r^2$ in the second term.

[7]The set of points of increase of a distribution function F is $\{r : F(r-\epsilon) < F(r+\epsilon)\ \ \forall \epsilon > 0\}$.

We prove the result by contradiction. Hence, we now assume that the optimal input distribution F_0 has an infinite number of points of increase on a bounded interval. Next, we extend the $\Phi(\cdot)$ in (70) to the complex domain:

$$\Phi(z) = E_{\mathsf{K}} \left\{ \int_0^\infty f_{R|r,\mathsf{K}}(R|z,\mathsf{K}) \log\, g(R,F_r,\mathsf{K})\, dR \right\} + \log(1+z^2) + C + (m-1) \tag{76}$$

where $z \in \mathbb{C}$ and $\log$ is the principle branch of the logarithm. The Identity Theorem for analytic functions [41] states that if two functions are analytic in a region and if they coincide for an infinite number of distinct points having a limiting point, they are equal everywhere in that region. It is shown in Appendix C that $\Phi(z)$ is analytic in a region $\mathcal{D}$ that includes the positive real line. By the above assumption on the optimal input distribution, $\Phi(z) = 0$ for an infinite number of points having a limiting point[8] in region $\mathcal{D}$. Therefore, by the Identity Theorem, we should have $\Phi(r) = 0$ for all $r \geq 0$. Clearly, this is not possible from the upper bound in (74) which diverges to $-\infty$ as $r \to \infty$. Hence, the optimal input cannot have an infinite number of points of increase on a bounded interval, from which we conclude that the optimal input distribution is discrete with a finite number of mass points. □

After the characterization of the discrete nature of the optimal input, the optimization problem in (69) can be solved using vector optimization techniques. Numerical results indicate that the optimal magnitude distribution F_r has a single mass at the peak level $r = \sqrt{L}$ for low-to-medium received peak SNR $= \frac{\gamma^2 P}{N_0}$ levels. Hence, all the information is carried by the isotropically distributed directional unit vector. Therefore, information transmission is achieved by sending points on the surface of an $(m-1)$-dimensional complex sphere with radius $\frac{\sqrt{LN_0}}{\tilde{\gamma}}$. Note that the mutual information (in nats per m symbols) achieved by having a single-mass at $r = \sqrt{L}$ is

$$I_{cm} = -E_{\mathsf{K}} \left\{ \int_0^\infty f_{R|r,\mathsf{K}}(R|r=\sqrt{L},\mathsf{K}) \log\, g(R,F_r,\mathsf{K})\, dR \right\} - \log(1+L) - (m-1). \tag{77}$$

Figure 6 plots the capacity values as a function of SNR for block lengths of $m = 10, 20, 30$ and 40. These capacity values are achieved with optimal power allocation. The optimal fractions of power allocated to the pilot symbol are plotted in Fig. 7. Note that for the range of SNR values considered in the figure, the optimal value of δ is slightly smaller than $1/m$ and approaches $1/m$ as SNR tends to 0. This power allocation strategy is significantly different from that of the worst-case scenario in which $\delta^* \to 1/2$ with decreasing SNR.

In the low-SNR regime, the tradeoff between spectral efficiency and energy per bit obtained from $\frac{E_b}{N_0} = \frac{\text{SNR}\log 2}{C(\text{SNR})}$ is the key performance measure [14]. If we assume, without loss of generality, that one symbol occupies a 1s$\times$1Hz time-frequency slot, then the maximum spectral efficiency is $\mathsf{C}(E_b/N_0) = C(\text{SNR})\log_2 e$ bits/s/Hz where we have assumed that $C(\text{SNR})$ is in nats/symbol. Fig. 8 plots the bit energy values as a function of the spectral efficiency. It is again observed that the minimum bit energy is achieved at a nonzero spectral efficiency and the required

[8]The Bolzano-Weierstrass Theorem [40] states that every bounded infinite set of real numbers has a limit point.

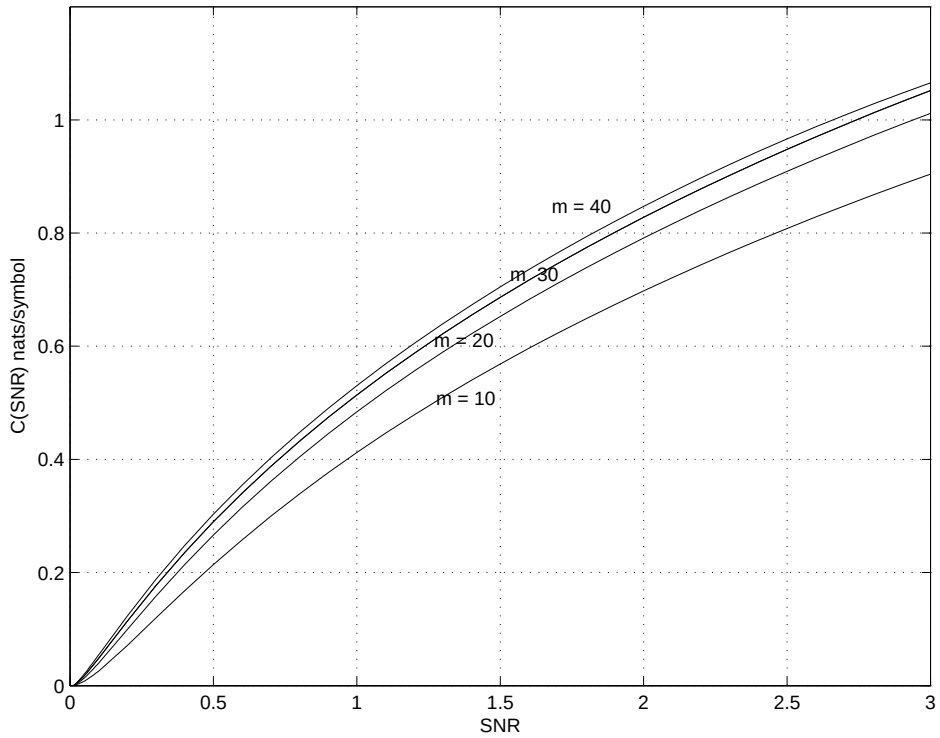

Fig. 6. Capacity (nats/symbol) vs. SNR for block lengths of $m = 10, 20, 30$ and 40 when the input is subject to peak power limitations.

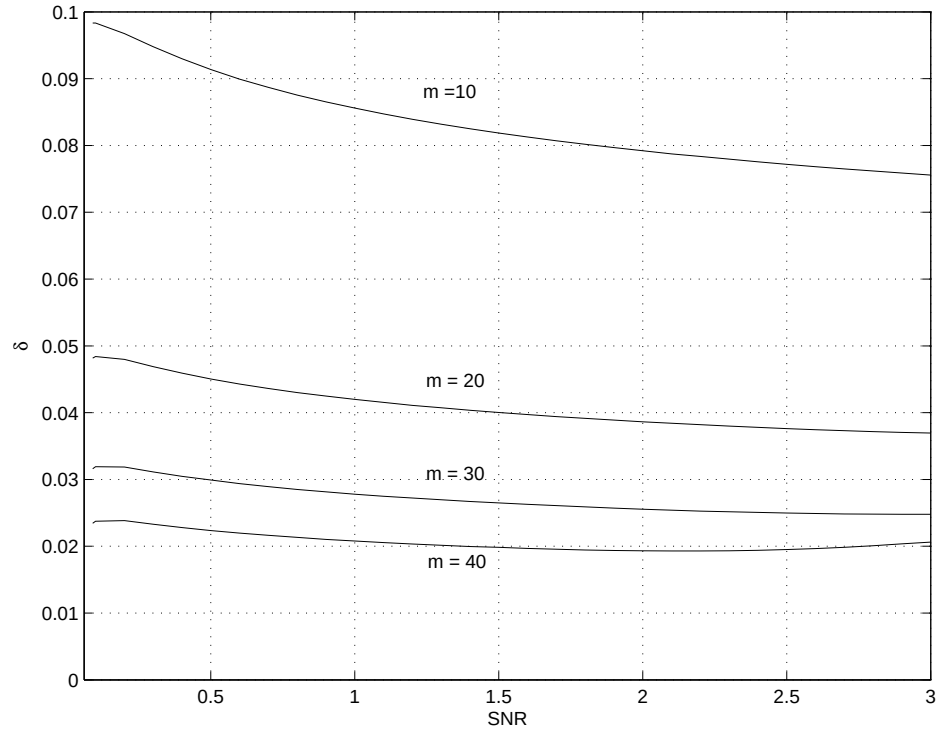

Fig. 7. Optimal fraction of power δ allocated to the pilot symbol vs. SNR for block lengths of $m = 10, 20, 30$ and 40.

bit energy values grow without bound as SNR and hence the spectral efficiency is further decreased. Indeed, we can show the following result.

Theorem 7: Assume that the normalized input magnitude distribution has a single mass and hence the magnitude is fixed at $r = \sqrt{L}$. For any value of $\delta \in (0, 1)$, the normalized bit energy required by this input grows without bound as the signal-to-noise ratio decreases to zero, i.e.,

$$\left.\frac{E_{b,cm}}{N_0}\right|_{I_{cm}=0} = \lim_{\text{SNR}\to 0} \frac{m\text{SNR}}{I_{cm}(\text{SNR})} \log 2 = \frac{m \log 2}{\dot{I}_{cm}(0)} = \infty. \tag{78}$$

Proof: Recall that $L = \frac{(1-\delta)m\text{SNR}}{\delta m\text{SNR}+1}$ and $\text{SNR} = \frac{\gamma^2 P}{N_0}$. Also, note that an expression for I_{cm} is given in (77). By making a change of variables, we have the following equivalent expression:

$$I_{cm} = -E_{\mathsf{K}}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|\sqrt{L}, \mathsf{K}\delta m\text{SNR}) \log\, g(R, F_r, \mathsf{K}\delta m\text{SNR})\, dR\right\} - \log(1+L) - (m-1) \tag{79}$$

where K is now an exponential random variable with $E\{\mathsf{K}\} = 1$, and hence is independent of SNR. We can easily show that

$$\left.\frac{\partial}{\partial \text{SNR}} f_{R|r,\mathsf{K}}(R|\sqrt{L}, \mathsf{K}\delta m\text{SNR})\right|_{\text{SNR}=0} = -(1-\delta)m\frac{R^{m-2}}{(m-2)!}\, e^{-R} + (1-\delta)m\frac{R^{m-1}}{(m-1)!}\, e^{-R}. \tag{80}$$

Note that

$$g(R, F_r, \mathsf{K}\delta m\text{SNR}) = \frac{(m-2)!}{R^{m-2}} f_{R|r,\mathsf{K}}(R|\sqrt{L}, \mathsf{K}\delta m\text{SNR}).$$

Using these facts, we can easily prove that

$$\dot{I}_{cm}(0) = \left.\frac{\partial I_{cm}}{\partial \text{SNR}}\right|_{\text{SNR}=0} = 0. \tag{81}$$

□

In the very low SNR regime, the channel estimate deteriorates and the performance approaches that of non-coherent Rayleigh block fading channels. As shown in [18], bit energy values required in these channels grow without bound as $\text{SNR} \to 0$ and the same phenomenon is observed here as well. In the worst-case scenario treated in Section IV, the performance deterioration at very low SNR levels is due to the fact that poor channel estimates are assumed to be perfect. In this section, similar observations are the result of the limitations on the peakedness of the signal. Nevertheless, designing the transmission and reception for channel in (58) rather than that in (11) leads to energy gains in the low-SNR regime. Fig. 9 provides a comparison of the bit energy values required in the worst case scenario and the scenario where peak power constraints are imposed and optimal signaling and decoding is employed. In the worst-case scenario, the channel estimate is assumed to be perfect and transmission and reception is designed for a known channel. This is obviously a poor assumption in the low-SNR regime and in Fig. 9 we observe bit energy gains of approximately 1.5 dB when optimal techniques are employed in the case of $m = 10$. Note that these gains are achieved when the input is subject to more stringent peak power constraints. From Fig. 9, we also conclude that in the low-SNR regime, the achievable rate expression in (15) is

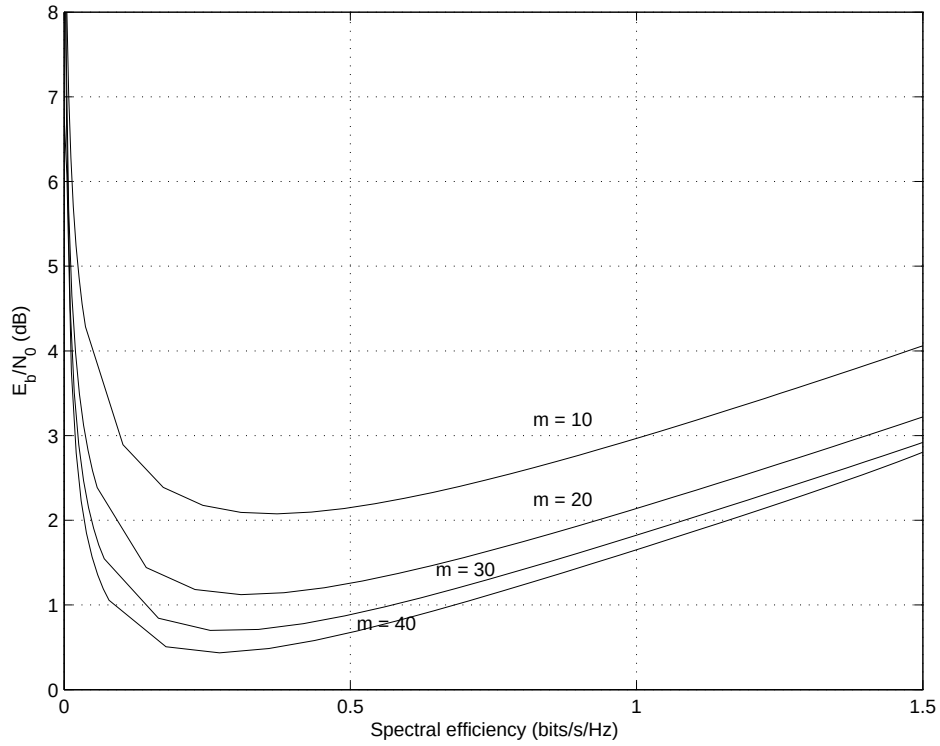

Fig. 8. Bit energy $\frac{E_b}{N_0}$ vs. Spectral efficiency $\mathsf{C}\left(\frac{E_b}{N_0}\right)$ in pilot-assisted systems with block lengths $m = 10, 20, 30$ and 40.

a lower bound to the peak-power limited capacity of the channel in (58). Note that (15) will eventually exceed this capacity value at high SNR levels as it is obtained under less strict average power constraints.

In training-based systems, certain fraction of time and power which otherwise will be used for data transmission is allocated to the pilot symbols to facilitate channel estimation. Hence, there is a potential for performance loss in terms of data rates. However, at the same time, the availability of channel estimates at the receiver tends to improve the performance. On the other hand, in noncoherent communications, there is no attempt for channel estimation and communication is performed over unknown channels. The analysis presented in this paper can be applied to noncoherent communications in a straightforward manner by choosing $\delta = 0$ and replacing m in the equations by $m+1$ as no time is allocated to pilot symbols. Hence, for instance, the discrete nature of the optimal input under peak power constraints can easily be shown for the noncoherent Rayleigh channel as well. However, the details of this analysis is omitted because the discreteness results are proven for noncoherent Rician fading channels in [18] and for more general noncoherent MIMO channels in [8]. Here, we present numerical results. Figures 10 and 11 compare the performances of training-based and noncoherent communication systems. In Fig. 10, the bit energy values are plotted for both schemes when the block length is $m = 20$. It is observed that for this relatively small value of the block length, both schemes achieve almost the same minimum bit energy value,

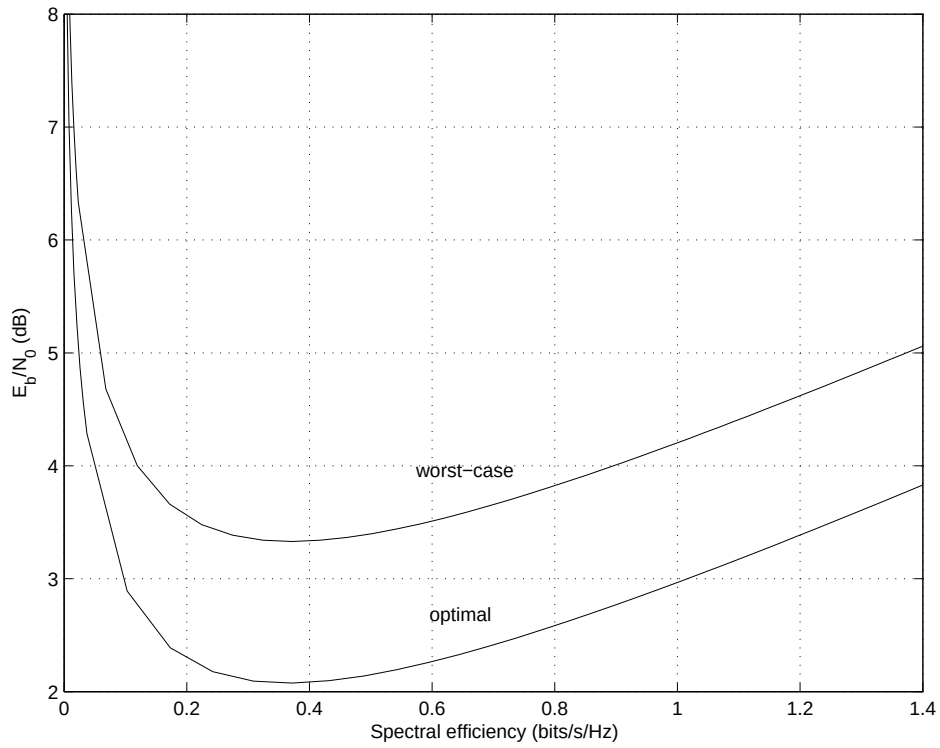

Fig. 9. Bit energy $\frac{E_b}{N_0}$ vs. Spectral efficiency $\mathsf{C}\left(\frac{E_b}{N_0}\right)$ in the worst-case scenario and the scenario of optimal coding-decoding under input peak-power constraints. The block length is $m = 10$.

and therefore, the training-based performance is surprisingly rather close to that of the noncoherent scheme even in the low-SNR regime. Fig. 11 plots the capacity values as a function of the block length at $\text{SNR} = 5$ dB. Here, we also observe that the performance of training-based schemes comes very close to that of the noncoherent scheme. Therefore, if having the channel estimate reduces the complexity of the receiver and/or pilot signals are additionally used for timing and frequency-offset synchronization or channel equalization, training-based schemes can be preferred over noncoherent communications with small loss in data rates.

A. Capacity with Ideal Interleaving and Per-symbol Peak Power Constraints

Since most of the well-known codes are designed to correct errors that occur independently from the location of other errors [43], practical communication systems employ interleavers at the transmitters to gain protection against error bursts. Deinterleavers are used at the receiver to reverse the interleaving operation. In this section, we consider such systems and assume that ideal interleaving is used so that each data symbol experiences independent channel conditions. Pilot symbols are inserted periodically after the interleaver. We note that a pilot-assisted transmission with ideal interleaving is also studied in [24] and [25] where achievable rates are considered.

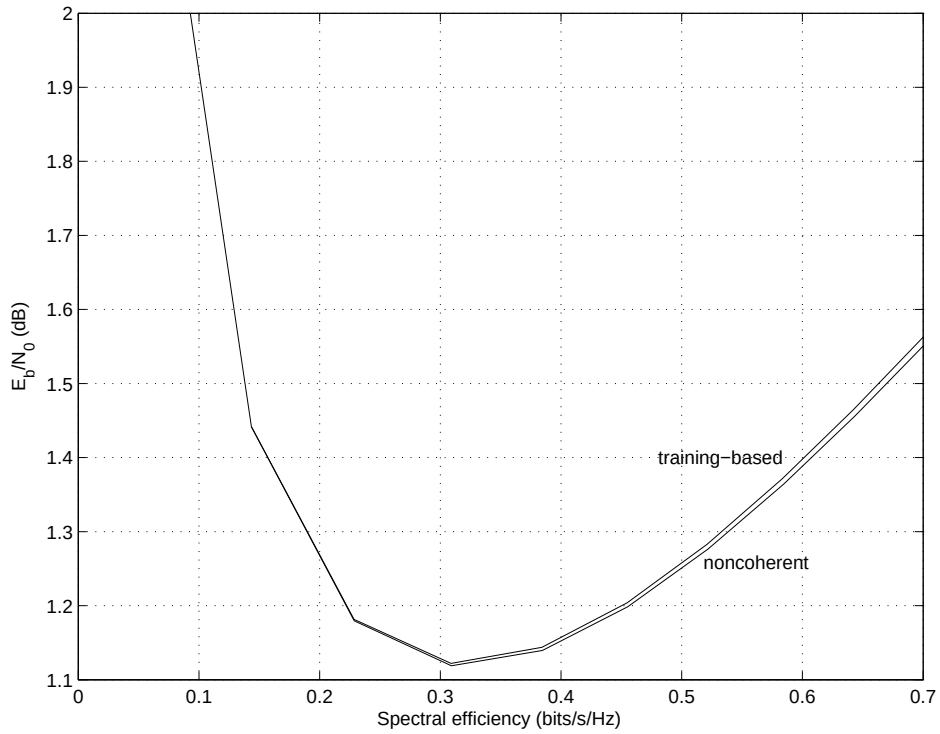

Fig. 10. Bit energy $\frac{E_b}{N_0}$ vs. Spectral efficiency $\mathsf{C}\left(\frac{E_b}{N_0}\right)$ for training-based and noncoherent communication systems when $m = 20$.

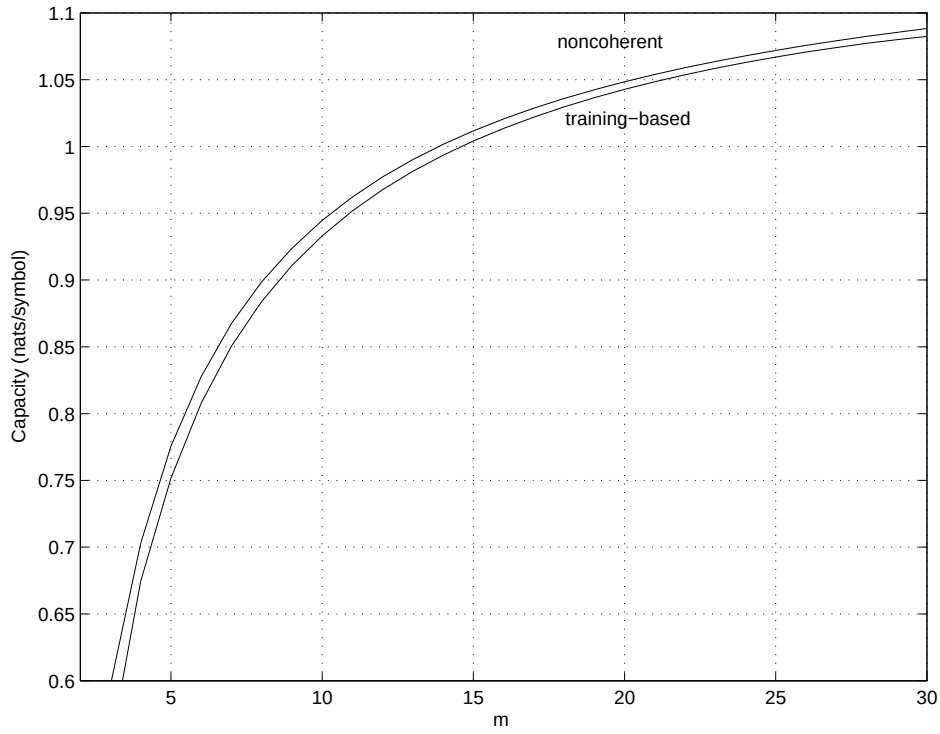

Fig. 11. Capacity (nats/symbol) vs. block length m for training-based and noncoherent communication systems. SNR $= 5$ dB.

Since interleaving breaks the channel correlation seen by the data symbols, channel memory can no longer be taken advantage of in the transmission. Hence, interleaving in general decreases the capacity. Therefore, the capacity results in this section can also be regarded as lower bounds on the capacity of a non-interleaved system. On the other hand, one advantage of interleaving is the simplification of signaling schemes.

We continue considering the block fading channel model. Hence, the channel stays constant for a block of m symbols. However, after deinterleaving, the channel output can be expressed as

$$y_{d,i} = \hat{h}_i x_{d,i} + \tilde{h}_i x_{d,i} + n_i \quad i = 1, 2, 3 \dots \tag{82}$$

Note that due to interleaving, each data symbol $x_{d,i}$ is affected by independent and identically distributed fading coefficients $h_i = \hat{h}_i + \tilde{h}_i$. In this section, we consider per-symbol peak power constraints, $|x_i|^2 \overset{a.s.}{\leq} P \; \forall i$. Therefore, the pilot symbol power is $|x_t|^2 = P$. Note that the use of more than one pilot may be optimal. The channel capacity in this setting is formulated as follows:

$$C = \sup_{1 \leq l \leq m} \sup_{\substack{x_d \\ |x_d|^2 \overset{a.s.}{\leq} P}} \frac{m-l}{m} I(x_d; y_d|\hat{h}) \tag{83}$$

where l denotes the number of pilot symbols per m symbols, and

$$\hat{h} \sim \mathcal{CN}\left(0, \frac{\gamma^4 l P}{\gamma^2 l P + N_0}\right) \text{ and } \tilde{h} \sim \mathcal{CN}\left(0, \frac{\gamma^2 N_0}{\gamma^2 l P + N_0}\right).$$

The inner maximization in (83) becomes a special case of the inner maximization in (62) when we reduce the dimensionality of the optimization problem in (62)[9] by choosing $m = 2$. Therefore, the results on the structure of the capacity-achieving input immediately apply to the setting we consider in this section. The optimal input has a uniformly distributed phase. With this characterization, the capacity is

$$C = \sup_{1 \leq l \leq m} \sup_{\substack{F_r \\ r \overset{a.s.}{\leq} \sqrt{L}}} \frac{m-l}{m} I(F_r|\hat{h}) \tag{84}$$

where

$$I(F_r|\hat{h}) = -E_{\mathsf{K},r}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) \log\, g(R, F_r, \mathsf{K})\, dR\right\} - E_r\{\log(1+r^2)\} - 1 \tag{85}$$

where

$$f_{R|r,\mathsf{K}}(R|r,\mathsf{K}) = \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{1+r^2} I_0\left(\frac{2\sqrt{\mathsf{K}R}\, r}{1+r^2}\right), \tag{86}$$

$$g(R, F_r, \mathsf{K}) = \int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K})\, dF_r, \tag{87}$$

[9]Note that the input constraints, error variances, and the constants multiplying the mutual information expressions will be different in the specialized case of (62) and in (83). But, the general structures of the two optimization problems are the same.

and, $R = \frac{|y_d|^2}{N_0}$, $r = \frac{\tilde{\gamma}|x_d|}{\sqrt{N_0}}$, $\mathsf{K} = \frac{|\hat{h}|^2}{\tilde{\gamma}^2}$, $\tilde{\gamma}^2 = \frac{\gamma^2 N_0}{\gamma^2 l P + N_0}$, and $L = \frac{\gamma^2 P/N_0}{l\gamma^2 P/N_0 + 1} = \frac{\text{SNR}}{l\text{SNR}+1}$. Note that K is an exponential random variable with mean $E\{\mathsf{K}\} = \frac{E\{|\hat{h}^2|\}}{\tilde{\gamma}^2} = \frac{l\gamma^2 P}{N_0} = l\text{SNR}$. Since the inner maximization in (84) is a special case of that in (69), we immediately have the following result.

Theorem 8: Fix the value of $1 \leq l \leq m$. The input distribution that maximizes the mutual information $I(F_r|\hat{h})$ in (84) is discrete with a finite number of mass points.

Next, we present numerical results. Fig. 12 plots, for different values of the block lengths, the capacity curves as a function of SNR for training-based schemes. We observe that the capacity values increase with the block length even though the channel in (82) is memoryless. This performance gain should be attributed to the fact that the channel estimate improves with increasing block length. Fig. 12 also plots the capacity of the interleaved noncoherent communications in which no attempt is made to learn the channel. From the comparison of the capacity curves, we observe that training significantly enhances the data rates when data symbols are interleaved at the transmitter. In Fig. 13, bit energy curves as a function of the spectral efficiency are plotted. Again, we see that training-based schemes perform much better in terms of energy efficiency than the noncoherent scheme. In all cases, the minimum bit energy is achieved at a nonzero spectral efficiency level below which one should not operate. The bit energy requirement increases without bound as spectral efficiency decreases to zero. When we compare Figs. 8 and 13, we note that while simplifying the system design, interleaving also incurs a penalty in energy efficiency. Finally, in Fig. 14, we provide the optimal resource allocations by plotting the optimal number of pilot symbols per block as a function of SNR for different block length values. We realize that optimal number of pilots tends to increase as SNR decreases and approaches $m/2$. Hence, as in Section IV-A, asymptotically half of the available power in each block should be allocated to the training symbols.

B. Achievable Rates and Bit Energies of On-Off Keying

In this section, we relax the input constraints and assume that the input is subject to an average power constraint

$$E\{\|\mathbf{x}\|^2\} \leq mP. \tag{88}$$

We consider the channel model (58) where there is no interleaving. Akin to Section IV-C, our goal is to obtain the attainable bit energy levels when signals with high peak-to-average power ratios are employed. As before, single pilot symbol with power $|x_t|^2 = \delta m P$ is used and hence the data vector is subject to $E\{\|\mathbf{x}_d\|^2\} \leq (1-\delta)mP$. The data vector is again assumed to have an isotropically distributed directional vector $\mathbf{v}$, and hence $\mathbf{x}_d = \|\mathbf{x}_d\|\mathbf{v}$. We further assume that the on-off keying is used for magnitude modulation and therefore

$$\frac{1}{\sqrt{m}}\|\mathbf{x}_d\| = \begin{cases} A & \text{with prob.} \quad p_0 \\ 0 & \text{with prob.} \quad 1-p_0 \end{cases} \tag{89}$$

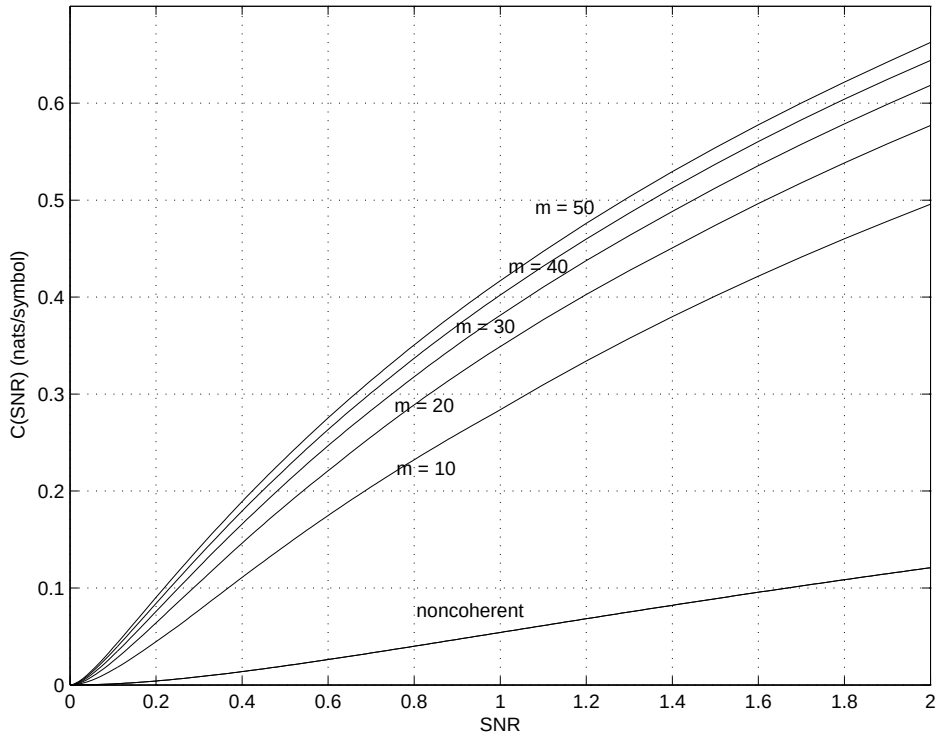

Fig. 12. Capacity (nats/symbol) vs. SNR for interleaved, training-based transmissions when block lengths are $m = 10, 20, 30, 40$ and 50, and for interleaved noncoherent transmission over the unknown Rayleigh fading channel.

where A is a fixed magnitude level that does not vary with the power P. In order to satisfy the average power constraint we should have

$$A^2 p_0 = (1-\delta)P \Rightarrow p_0 = \frac{(1-\delta)P}{A^2} \tag{90}$$

Therefore, in this signaling scheme, the peak power of the transmitted data signal is kept constant while its probability vanishes as $P \to 0$. Hence, while the peak power is fixed, the peak-to-average power ratio grows without bound as $P \to 0$. Similarly as before, we define $r = \frac{\tilde{\gamma}\|\mathbf{x}_d\|}{\sqrt{N_0}}$. With this definition, the distribution of r is

$$r = \begin{cases} \frac{\gamma\sqrt{m}A}{\sqrt{\gamma^2\delta m P + N_0}} & \text{with prob.} \quad p_0 = \frac{(1-\delta)P}{A^2} \\ 0 & \text{with prob.} \quad 1-p_0 \end{cases} . \tag{91}$$

We further define $\nu = \frac{A^2\gamma^2}{(1-\delta)N_0}$ which does not depend on P, and $\text{SNR} = \frac{\gamma^2 P}{N_0}$. Now, we can write

$$r = \begin{cases} r_0 = \frac{\sqrt{(1-\delta)m\nu}}{\sqrt{\delta m \text{SNR}+1}} & \text{with prob.} \quad p_0 = \frac{\text{SNR}}{\nu} \\ 0 & \text{with prob.} \quad 1-p_0 \end{cases} . \tag{92}$$

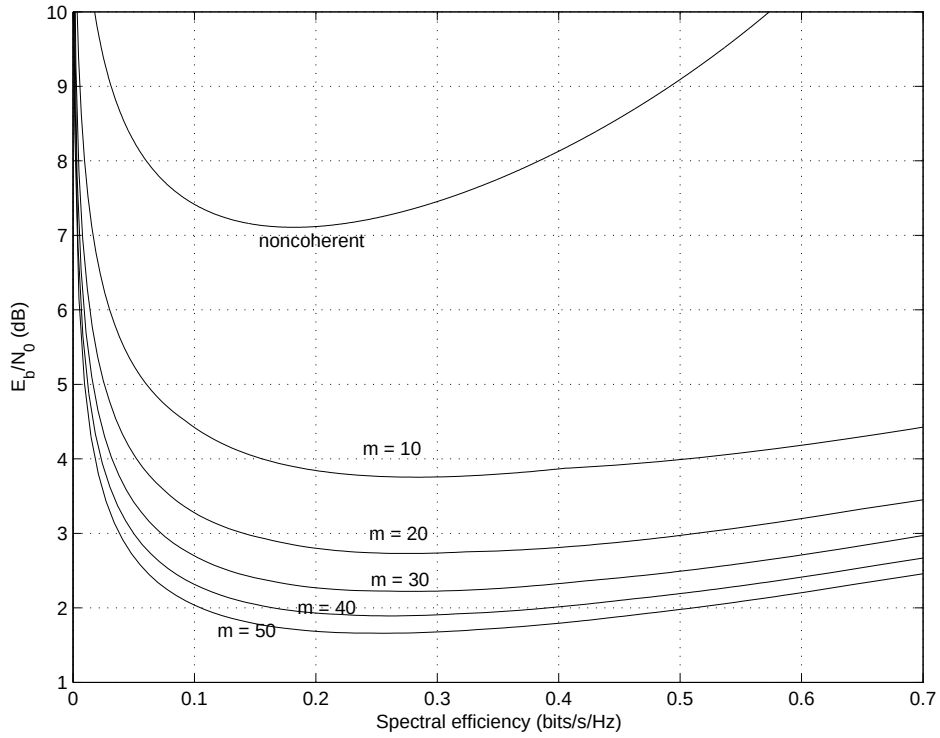

Fig. 13. Bit energy $\frac{E_b}{N_0}$ vs. Spectral efficiency $\mathsf{C}\left(\frac{E_b}{N_0}\right)$ for interleaved, training-based transmissions when block lengths are $m = 10, 20, 30, 40$ and 50, and for interleaved noncoherent transmission over the unknown Rayleigh fading channel.

For a given value of δ, the mutual information achieved by the isotropically distributed directional vector $\mathbf{v}$ and r whose distribution is given in (92) is

$$I_{ook} = -E_{\mathsf{K}}\left\{\int_0^\infty f_{R|\mathsf{K}}(R|\mathsf{K})\log\left(\frac{(m-2)!}{R^{m-2}}f_{R|\mathsf{K}}(R|\mathsf{K})\right)\,dR\right\} - p_0\log(1+r_0^2) - (m-1) \tag{93}$$

where $f_{R|\mathsf{K}}(R|\mathsf{K}) = (1-p_0)f_{R|r,\mathsf{K}}(R|r=0,\mathsf{K}) + p_0 f_{R|r,\mathsf{K}}(R|r=r_0,\mathsf{K})$ and $f_{R|r,\mathsf{K}}(R|r,\mathsf{K})$ is given in (67). Note that K is an exponential random variable with mean $E\{\mathsf{K}\} = \frac{\delta m\gamma^2 P}{N_0} = \delta m\text{SNR}$. Next, we obtain the bit energy required for reliable communications with OOK as $\text{SNR} \to 0$.

Theorem 9: Assume that the normalized input magnitude distribution is given by (92). For a given value of $\delta \in (0,1)$, the normalized bit energy required by this input as $P \to 0$ is

$$\left.\frac{E_{b,ook}}{N_0}\right|_{I_{ook}=0} = \lim_{\text{SNR}\to 0}\frac{m\text{SNR}}{I_{ook}(\text{SNR})}\log 2 = \frac{m\log 2}{\dot{I}_{ook}(0)} = \frac{\log 2}{(1-\delta) - \frac{1}{m\nu}\log(1+(1-\delta)m\nu)}. \tag{94}$$

Proof: As in the proof of Theorem 7, we apply a change of variables and express the mutual information as

$$I_{ook} = -E_{\mathsf{K}}\left\{\int_0^\infty f_{R|\mathsf{K}}(R|\mathsf{K}\delta m\text{SNR})\log\left(\frac{(m-2)!}{R^{m-2}}f_{R|\mathsf{K}}(R|\mathsf{K}\delta m\text{SNR})\right)\,dR\right\} - p_0\log(1+r_0^2) - (m-1) \tag{95}$$

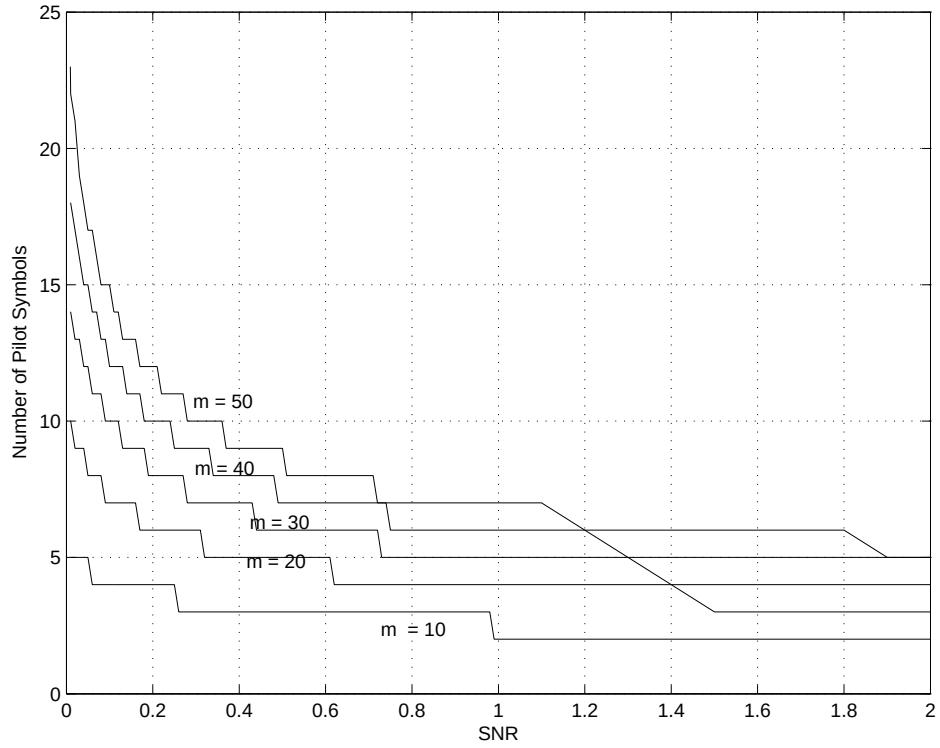

Fig. 14. Number of pilot symbols per block vs. SNR for interleaved, training-based transmissions when block lengths are $m = 10, 20, 30, 40$ and 50.

where K is now an exponential random variable with mean $E\{\mathsf{K}\} = 1$. It can be easily seen that

$$\frac{\partial}{\partial \text{SNR}} p_0 \log(1 + r_0^2)\bigg|_{\text{SNR}=0} = \frac{1}{\nu} \log(1 + (1-\delta)m\nu). \tag{96}$$

We can also show that

$$\frac{\partial}{\partial \text{SNR}} f_{R|\mathsf{K}}(R|\mathsf{K}\delta m \text{SNR})\bigg|_{\text{SNR}=0} = -\frac{1}{\nu} f_{R|r,\mathsf{K}}(R|r=0, K=0) + \frac{1}{\nu} f_{R|r,\mathsf{K}}(R|r=\sqrt{(1-\delta)m\nu}, K=0). \tag{97}$$

Using (97), we can prove that the derivative of the first term on the right-hand side of (95) with respect to SNR at $\text{SNR} = 0$ is $(1-\delta)m$. Combining this result with (96), we arrive to

$$\dot{I}_{ook}(0) = (1-\delta)m - \frac{1}{\nu} \log(1 + (1-\delta)m\nu) \tag{98}$$

which concludes the proof. □

Theorem 9 shows that unlike previously treated cases, reliable communications with OOK modulation with fixed peak power requires finite bit energy as $P \to 0$. Hence, OOK provides significant improvements in energy efficiency in the low-SNR regime at the cost of high peak-to-average power ratio. Since $\nu = \frac{A^2\gamma^2}{(1-\delta)N_0}$, we can also

express the asymptotic bit energy level as

$$\left.\frac{E_{b,ook}}{N_0}\right|_{I_{ook}=0} = \frac{\log 2}{(1-\delta)\left(1-\frac{1}{m\frac{A^2\gamma^2}{N_0}}\log\left(1+m\frac{A^2\gamma^2}{N_0}\right)\right)}. \tag{99}$$

It has been shown in [18] that, if noncoherent communications with no channel estimation is performed and the input is subject to $E\{\|\mathbf{x}|^2\} \leq mP$ and $\|\mathbf{x}\|^2 \overset{\text{a.s.}}{\leq} mA$, then optimal signaling requires the following bit energy value as $P \to 0$:

$$\left.\frac{E_{b,noncoh}}{N_0}\right|_{\mathsf{C}=0} = \frac{\log 2}{\left(1-\frac{1}{m\frac{A^2\gamma^2}{N_0}}\log\left(1+m\frac{A^2\gamma^2}{N_0}\right)\right)}. \tag{100}$$

We note that similar results for fading channels with memory are obtained in [21] through the analysis of capacity per unit cost. Comparing (99) and (100), we find that training-based schemes suffer an energy penalty due to the presence of the term $1/(1-\delta)$ and this penalty vanishes if $\delta \to 0$. Therefore, if OOK with fixed power is employed, the power of the training symbols should be decreased to zero as $P \to 0$ to match the noncoherent performance. This power allocation policy is in stark contrast to the results in the previous sections. Note that as SNR decreases, data transmission occurs extremely infrequently. In such a case, performing channel estimation all the time for each m-block irrespective of whether or not data transmission takes place is not an good design choice. Hence, a gradual decrease in the power allocated to training should also be intuitively expected. We further remark that as $m \to \infty$ and $\delta \to 0$, $\left.\frac{E_{b,ook}}{N_0}\right|_{I_{ook}=0} \to -1.59$ dB.

Fig. 15 plots the bit energy levels as a function of spectral efficiency for training-based OOK with fixed peak power and for training-based optimal signaling under input peak power constraints in the form $\|\mathbf{x}\|^2 \overset{\text{a.s.}}{\leq} (1-\delta)mP$. In this figure, the block length is $m = 10$, and for OOK, $\nu = 1$. As predicted, below the spectral efficiency of approximately 0.4 bits/s/Hz, OOK provides better energy efficiency. The bit energy requirements of OOK decreases as spectral efficiency decreases as opposed to the behavior presented in the peak-power-limited case. Numerical analysis have also shown that the fraction of power allocated to training, δ, in OOK decreases as SNR decreases, conforming with the discussion in the previous paragraph.

VI. Conclusion

In this paper, we have studied the energy efficiency and capacity of training-based communication schemes employed for the transmission of information over a-priori unknown Rayleigh block fading channels. We have initially considered the worst-case scenario in which the product of the estimate error and transmitted signal is assumed to be Gaussian noise. The capacity expression obtained under this assumption is a lower bound to the true capacity of the channel, and provides the achievable rates when the communication system is designed as if the channel estimate were perfect. We have investigated the bit energy levels required for reliable communications and

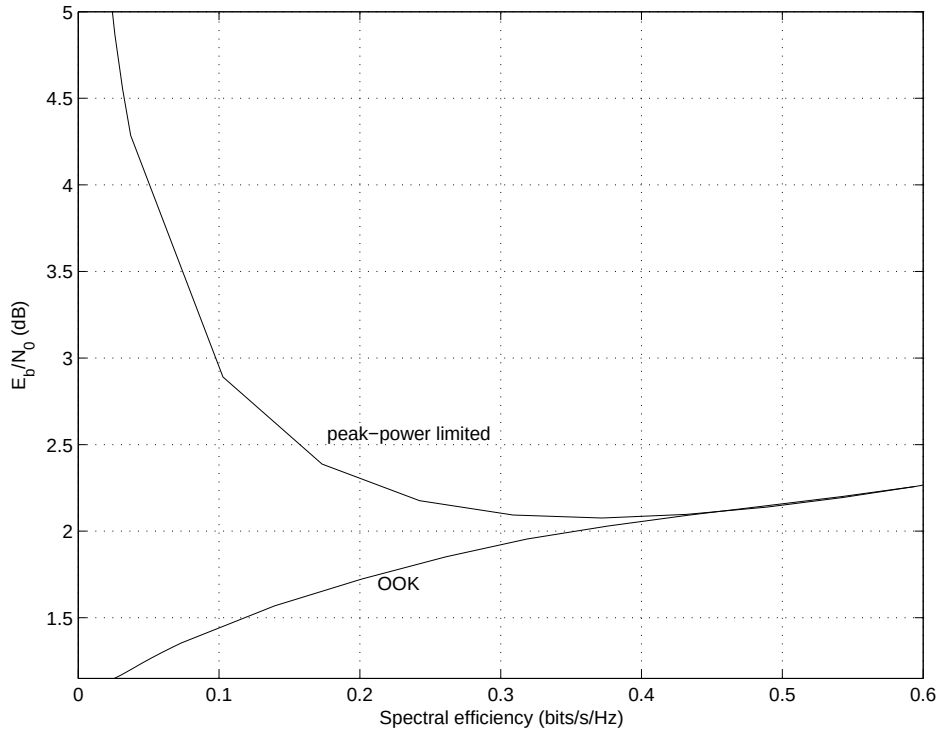

Fig. 15. Bit energy $\frac{E_b}{N_0}$ vs. Spectral efficiency $\mathsf{C}\left(\frac{E_b}{N_0}\right)$ for training-based OOK signaling and training-based optimal signaling under input peak power constraints. The block length is $m = 10$.

quantified the penalty in energy efficiency incurred due to regarding the imperfect channel estimate as perfect in the low-SNR regime. We have shown that the bit energy requirements grow without bound as $\text{SNR} \to 0$ regardless of the size of the block length m. Hence, the minimum bit energy is achieved at a nonzero SNR value below which one should not operate under the aforementioned assumptions. We have also shown that approaching the minimum bit energy level of -1.59 dB is extremely slow in terms of block length as $m \to \infty$. Similar results are obtained if peak power limitations are imposed on training symbols. We have also investigated flash training and transmission schemes to improve the energy efficiency at low SNR levels. We have shown that in order for the bit energy requirement not to grow as $\text{SNR} \to 0$, the duty cycle in flash transmission should vanish linearly with decreasing SNR.

Next, we have analyzed the capacity and energy efficiency of training-based schemes when the input is subject to peak power constraints. We have characterized that the capacity-achieving input has a discrete magnitude and an isotropically distributed unit directional vector. Using this characterization, we have obtained the capacity expressions, optimal training power allocations, and bit energy levels required for reliable communications. We have noted that at low SNRs, the optimal input magnitude is fixed at a constant level. Due to the presence of

the peak power constraints, the bit energy requirements are again shown to increase without bound as $\text{SNR} \to 0$. However, we have seen that gains in energy efficiency are obtained when optimal signaling and decoding are employed. We have compared the performances of training-based and noncoherent transmission schemes. Although training-based schemes dedicate certain amount of time and power to training symbols and as a result are expected to suffer in terms of data rates, we have observed that the performance loss is small even at relatively small block lengths and small SNR levels. We have also considered the case in which interleaving used at the transmitter for protection against error bursts and per-symbol peak power constraints are imposed. We have obtained the channel capacity, optimal training duration, and analyzed the energy efficiency. In this case, training is shown to improve the performance with respect to noncoherent communications. We have also investigated the improvements in energy efficiency in the low-SNR regime if OOK with fixed peak power and vanishing duty cycle is employed at the transmitter. Finally, we note that this work has primarily focused on block fading channels. Recently, we in [33], [34] and [35] have considered more general fading processes with memory. Since the exact capacity is rather difficult to obtain in such cases, achievable rate expressions are analyzed, and subsequently energy efficiency and optimal resource allocations are studied.

APPENDIX

A. Derivation of the Mutual Information Expression in Theorem 4

The input-output mutual information expression for channel (58) is

$$I(\mathbf{x}_d;\mathbf{y}_d|\hat{h}) = E_{\hat{h}}E_{\mathbf{x}_d}\int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})\log\frac{f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})}{f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})}\,d\mathbf{y} \tag{101}$$

$$= -E_{\hat{h}}E_{\mathbf{x}_d}\int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})\log f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})\,d\mathbf{y} - E_{\mathbf{x}_d}\{\log(\pi^{m-1}N_0^{m-2}e^{m-1}(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0))\} \tag{102}$$

$$= -E_{\hat{h}}E_{\mathbf{x}_d}\int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})\log f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})\,d\mathbf{y} - E_r\{\log(1+r^2)\} - \log(\pi^{m-1}N_0^{m-1}) - (m-1) \tag{103}$$

Note that the second part of (102) is the conditional differential entropy of $\mathbf{y}$ given $\mathbf{x}_d$ and $\hat{h}$. (103) follows from the definition $r = \frac{\tilde{\gamma}\|x_d\|}{\sqrt{N_0}}$. The main difficulty is to simplify

$$\chi(\mathbf{x}_d,\hat{h}) = \int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})\log f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})\,d\mathbf{y} \tag{104}$$

which, in general, is an $2(m-1)$-fold integral. Note that

$$f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h}) = \int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})\,dF_{\mathbf{x}_d}. \tag{105}$$

Using the facts that $f_{\mathbf{y}|\mathbf{x}}(\Phi\mathbf{y}|\Phi\mathbf{x}_d,\hat{h}) = f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h})$ and input has circular symmetry, we can easily see that for any fixed unitary matrix Φ

$$f_{\mathbf{y}|\hat{h}}(\Phi\mathbf{y}|\hat{h}) = f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h}) = f_{\mathbf{y}|\hat{h}}(\|\mathbf{y}\||\hat{h}) \tag{106}$$

and hence

$$\chi(\Phi\mathbf{x}_d,\hat{h}) = \chi(\mathbf{x}_d,\hat{h}) = \chi(\|\mathbf{x}_d\|,\hat{h}). \tag{107}$$

Therefore, $f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})$ and $\chi(\mathbf{x}_d,\hat{h})$ are circularly-symmetric functions depending only on $\|\mathbf{y}\|$ and $\|\mathbf{x}_d\|$, respectively. Noting that

$$(\tilde{\gamma}^2\mathbf{x}_d\mathbf{x}_d^\dagger + N_0 I)^{-1} = \frac{\mathbf{I}}{N_0} - \frac{\tilde{\gamma}^2\mathbf{x}_d\mathbf{x}_d^\dagger}{N_0(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)}, \tag{108}$$

and defining $\mathbf{x}_d = \|\mathbf{x}_d\|\mathbf{v}$ and $\mathbf{y} = \|\mathbf{y}\|\mathbf{w}$, we can, after some algebraic steps, rewrite the conditional density function in (64) as

$$f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d,\hat{h}) = \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{N_0} - \frac{|\hat{h}|^2\|\mathbf{x}_d\|^2}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0} + \frac{\tilde{\gamma}^2\|\mathbf{x}_d\|^2\|\mathbf{y}\|^2|\mathbf{w}^\dagger\mathbf{v}|^2}{N_0(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} + \frac{2\|\mathbf{x}_d\|\|\mathbf{y}\||\hat{h}|\,\Re(e^{j\theta_{\hat{h}}}\mathbf{w}^\dagger\mathbf{v})}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0}\right)}{\pi^{m-1}N_0^{m-2}(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} \tag{109}$$

where $\Re(z)$ denotes the real part of the complex number z, and $\theta_{\hat{h}}$ is the phase of $\hat{h}$. The usefulness of (109) comes from the property that the magnitude $\|\mathbf{x}_d\|$ and the directional unit vector $\mathbf{v}$ are separated. We know from Theorem 3 that $\mathbf{v}$ is isotropically distributed and independent of $\|\mathbf{x}_d\|$. Hence, we now have

$$f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h}) = \int \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{N_0} - \frac{|\hat{h}|^2\|\mathbf{x}_d\|^2}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0} + \frac{\tilde{\gamma}^2\|\mathbf{x}_d\|^2\|\mathbf{y}\|^2|\mathbf{w}^\dagger\mathbf{v}|^2}{N_0(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} + \frac{2\|\mathbf{x}_d\|\|\mathbf{y}\||\hat{h}|\,\Re(e^{j\theta_{\hat{h}}}\mathbf{w}^\dagger\mathbf{v})}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0}\right)}{\pi^{m-1}N_0^{m-2}(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} f_{\mathbf{v}}(\mathbf{v})dF_{\|\mathbf{x}_d\|}. \quad (110)$$

where $f_{\mathbf{v}}$ is the probability density function of $\mathbf{v}$. Since $f_{\mathbf{y}|\hat{h}}$ is a function of only $\|\mathbf{y}\|$, we can, without loss of generality, assume that $\mathbf{w}^\dagger = [1, 0, 0, \ldots, 0]$. In such a case,

$$f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h}) = \int \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{N_0} - \frac{|\hat{h}|^2\|\mathbf{x}_d\|^2}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0} + \frac{\tilde{\gamma}^2\|\mathbf{x}_d\|^2\|\mathbf{y}\|^2|v_1|^2}{N_0(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} + \frac{2\|\mathbf{x}_d\|\|\mathbf{y}\||\hat{h}|\,\Re(e^{j\theta_{\hat{h}}}v_1)}{\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0}\right)}{\pi^{m-1}N_0^{m-2}(\tilde{\gamma}^2\|\mathbf{x}_d\|^2+N_0)} f_{v_1}(v_1)dF_{\|\mathbf{x}_d\|}. \quad (111)$$

where v_1 is the first component of $\mathbf{v}$ and f_{v_1} is the corresponding density function. From [5], we have for $m \geq 3$

$$f_{v_1}(v_1) = \frac{1}{2\pi}2(m-2)(1-|v_1|^2)^{m-3} \quad |v_1| \leq 1. \quad (112)$$

Hence, v_1 has a uniform phase and a magnitude whose density function is

$$f_{|v_1|}(|v_1|) = 2(m-2)|v_1|(1-|v_1|^2)^{m-3}. \quad (113)$$

Note that if $m = 2$, then $\mathbf{x}_d$ is one-dimensional and hence $\mathbf{x}_d = \|\mathbf{x}_d\|\mathbf{v} = \|\mathbf{x}_d\|e^{\theta_{\mathbf{x}_d}}$. Therefore, in this case, $|\mathbf{v}| = |v_1| = 1$ with probability one. Using these facts and defining $r = \frac{\tilde{\gamma}\|\mathbf{x}_d\|}{\sqrt{N_0}}$, $R = \frac{\|\mathbf{y}\|^2}{N_0}$, $\mathsf{K} = \frac{|\hat{h}|^2}{\tilde{\gamma}^2}$, and $a = |v_1|^2$, we obtain

$$f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h}) = \begin{cases} \int_0^\infty dF_r \, \frac{(m-2)e^{-R-\frac{\mathsf{K}r^2}{1+r^2}}}{\pi^{m-1}N_0^{m-1}(1+r^2)} \int_0^1 (1-a)^{m-3} e^{\frac{ar^2R}{1+r^2}} I_0\left(\frac{2\sqrt{\mathsf{K}R}\,r\sqrt{a}}{1+r^2}\right) da & m \geq 3 \\ \int_0^\infty dF_r \, \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{\pi^{m-1}N_0^{m-1}(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}\,r}{1+r^2}\right) & m = 2 \end{cases} \quad (114)$$

$$= \frac{g(R, F_r, \mathsf{K})}{\pi^{m-1}N_0^{m-1}} \quad (115)$$

where $g(R, F_r, \mathsf{K})$ is defined in (68). Therefore, we have

$$\chi(\mathbf{x}_d, \hat{h}) = \int f_{\mathbf{y}|\mathbf{x}_d,\hat{h}}(\mathbf{y}|\mathbf{x}_d, \hat{h}) \log f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})\, d\mathbf{y} \quad (116)$$

$$= E_{\mathbf{y}|\mathbf{x}_d,\hat{h}}\{\log f_{\mathbf{y}|\hat{h}}(\mathbf{y}|\hat{h})\, d\mathbf{y}\} \quad (117)$$

$$= E_{R|r,\mathsf{K}}\{\log\left(\frac{g(R, F_r, \mathsf{K})}{\pi^{m-1}N_0^{m-1}}\right)\} \quad (118)$$

$$= -\log(\pi^{m-1}N_0^{m-1}) + E_{R|r,\mathsf{K}}\{\log g(R, F_r, \mathsf{K})\} \quad (119)$$

$$= -\log(\pi^{m-1}N_0^{m-1}) + \int_0^\infty f_{R|r,\mathsf{K}}(R|r, \mathsf{K}) \log g(R, F_r, \mathsf{K})\, dR \quad (120)$$

where $f_{R|r,\mathsf{K}}(R|r, \mathsf{K})$ is the conditional density function of R given r and K. Combining (103) and (120), we get

$$I(\mathbf{x}_d; \mathbf{y}_d|\hat{h}) = -E_{\mathsf{K},r}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r, \mathsf{K}) \log\, g(R, F_r, \mathsf{K})\, dR\right\} - E_r\{\log(1+r^2)\} - (m-1) \quad (121)$$

which is the mutual information expression provided in Theorem 4. Proof will be completed by showing that $f_{R|r,\mathsf{K}}(R|r,\mathsf{K})$ has the expression given in (67). From the previous development, we can easily verify that

$$f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R_1,\ldots,R_{m-1}|r,\mathsf{K}) = \begin{cases} \frac{(m-2)e^{-R-\frac{\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} \int_0^1 (1-a)^{m-3} e^{\frac{ar^2R}{1+r^2}} I_0\left(\frac{2\sqrt{\mathsf{K}R}r\sqrt{a}}{1+r^2}\right) da & m \geq 3 \\ \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}r}{1+r^2}\right) da & m = 2 \end{cases} \tag{122}$$

where $f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}$ is the conditional joint density function of $R_1,\ldots,R_{m-1}$ given r,K. Note that we above have defined $R_i = \frac{|y_i|^2}{N_0}$ and hence $R = \frac{\|\mathbf{y}\|^2}{N_0} = R_1 + R_2 + \ldots + R_{m-1}$. Note that the joint probability density function depends on the sum R. We have the following relationship

$$\int_0^\infty f_R(R|r,\mathsf{K})\, dR = \int f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R_1,\ldots,R_{m-1}|r,\mathsf{K})\, dR_1 \ldots dR_{m-1} \tag{123}$$

$$= \int dR_2 \ldots dR_{m-1} \int_{R_2+\ldots+R_{m-1}}^\infty f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R|r,\mathsf{K})\, dR \tag{124}$$

$$= \int dR_3 \ldots dR_{m-1} \int_{R_3+\ldots+R_{m-1}}^\infty f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R|r,\mathsf{K})\, dR \int_0^{R-(R_3+\ldots+R_{m-1})} dR_2 \tag{125}$$

$$= \int dR_3 \ldots dR_{m-1} \int_{R_3+\ldots+R_{m-1}}^\infty (R-(R_3+\ldots+R_{m-1})) f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R|r,\mathsf{K})\, dR \tag{126}$$

$$= \int_0^\infty \frac{R^{m-2}}{(m-2)!} f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R|r,\mathsf{K})\, dR. \tag{127}$$

(124) follows by applying the change of variables with $R = R_1 + R_2 + \ldots + R_{m-1}$ in the integral with respect to R_1. (125) is obtained by interchanging the integrals with respect to R_2 and R. (126) follows by evaluating the rightmost integral in (125). Finally, (127) is obtained through the repeated application of this procedure. From (127), we have

$$f_R(R|r,\mathsf{K}) = \frac{R^{m-2}}{(m-2)!} f_{R_1,\ldots,R_{m-1}|r,\mathsf{K}}(R|r,\mathsf{K}) \tag{128}$$

$$= \begin{cases} \frac{R^{m-2}}{(m-3)!} \frac{e^{-R-\frac{\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} \int_0^1 (1-a)^{m-3} e^{\frac{ar^2R}{1+r}} I_0\left(\frac{2\sqrt{\mathsf{K}R}r\sqrt{a}}{1+r^2}\right) da & m \geq 3 \\ \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}r}{1+r^2}\right) da & m = 2 \end{cases} \tag{129}$$

which is the same as the expression in (67).

B. Proof of Theorem 5

1) Existence of the Capacity-Achieving Input Distribution: An optimal distribution exists if the space of input distribution functions over which the minimization is performed is compact, and the objective functional is weak* continuous [39]. The compactness of the space of input distributions with second moment constraints is shown in [4]. The compactness for the more stringent case of peak limited inputs follows immediately from this result.

Therefore, we need only to show the weak* continuity of $I(\cdot|\hat{h})$. The weak* continuity of the functional $I(\cdot|\hat{h})$ is equivalent to

$$F_n \overset{w^*}{\to} F \Rightarrow I(F_n|\hat{h}) \to I(F|\hat{h}). \tag{130}$$

We first note the upper bound

$$f_R(R|r,\mathsf{K}) \leq \frac{R^{m-2}}{(m-3)!}\frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}\,r}{1+r^2}\right) \tag{131}$$

which is obtained from the bound $(1-a)^{m-3}e^{\frac{ar^2R}{1+r^2}}\, I_0\left(\frac{2\sqrt{\mathsf{K}R}\,r\sqrt{a}}{1+r^2}\right) \leq e^{\frac{r^2R}{1+r^2}}\, I_0\left(\frac{2\sqrt{\mathsf{K}R}\,r}{1+r^2}\right)\ \forall a \in [0,1]$. The upper bound in (131) is bounded for all $r \in [0,\sqrt{L}]$ and also for all $R \geq 0$ due to the exponential decrease in R in the second term. Since $f_R(R|r,\mathsf{K})$ and $\log(1+r^2)$ are continuous and bounded functions for all $r \in [0,\sqrt{L}]$ and $R \geq 0$, by the definition of weak convergence [39],

$$F_n \overset{w^*}{\to} F \Rightarrow \int_0^\infty \log(1+r^2)\,\mathrm{d}F_n(r) \to \int_0^\infty \log(1+r^2)\,\mathrm{d}F(r) \tag{132}$$

and

$$F_n \overset{w^*}{\to} F \Rightarrow \int_0^\infty f_R(R|r,\mathsf{K})\,\mathrm{d}F_n(r) \to \int_0^\infty f_R(R|r,\mathsf{K})\,\mathrm{d}F(r) \tag{133}$$

for all $R \geq 0$. Therefore, we have

$$F_n \overset{w^*}{\to} F \Rightarrow g(R,F_n,r) \to g(R,F,r) \quad \forall R \geq 0. \tag{134}$$

Note that the mutual information in (66) can also be written as

$$I(F_r|\hat{h}) = -\int_0^\infty d\mathsf{K}\, f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, g(R,F_r,\mathsf{K}) \log\, g(R,F_r,\mathsf{K}) - E_r\{\log(1+r^2)\} - (m-1) \tag{135}$$

The weak* continuity of the second term on the right-hand-side of (135) follows from (132). In order to show (130) and hence the weak* continuity of the mutual information, we need to prove

$$\lim_{n\to\infty} \int_0^\infty d\mathsf{K}\, f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, g(R,F_n,\mathsf{K}) \log\, g(R,F_n,\mathsf{K}) \tag{136}$$

$$= \int_0^\infty \lim_{n\to\infty} d\mathsf{K}\, f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, g(R,F_n,\mathsf{K}) \log\, g(R,F_n,\mathsf{K}) \tag{137}$$

$$= \int_0^\infty d\mathsf{K}\, f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty \lim_{n\to\infty} dR\, \frac{R^{m-2}}{(m-2)!}\, g(R,F_n,\mathsf{K}) \log\, g(R,F_n,\mathsf{K}) \tag{138}$$

$$= \int_0^\infty d\mathsf{K}\, f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, g(R,F,\mathsf{K}) \log\, g(R,F,\mathsf{K}) \tag{139}$$

(139) follows from (134) and the continuity of the function $x\log x$. In order to justify the interchanges of the limit and integral in (137) and (138), we invoke the Dominated Convergence Theorem [40] which requires an

integrable upper bound on the integrand. We first find the following upper bound on the function g:

$$g(R, F_n, \mathsf{K}) \leq (m-2) \int_0^{\sqrt{L}} \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}\, r}{1+r^2}\right) dF_n(r) \tag{140}$$

$$\leq (m-2) e^{-\frac{R}{1+L}+\sqrt{\mathsf{K}R}} \int_0^{\sqrt{L}} \frac{e^{-\frac{\mathsf{K}r^2}{1+r^2}}}{(1+r^2)}\, dF_n(r) \tag{141}$$

$$\leq (m-2)\, e^{-\frac{R}{1+L}+\sqrt{\mathsf{K}R}} \tag{142}$$

$$\triangleq u(R, \mathsf{K}) \quad \forall n, \quad \forall R, \mathsf{K} \geq 0. \tag{143}$$

(140) follows from the upper bound in (131). (141) is obtained by noting that $e^{-\frac{R}{1+r^2}} \leq e^{-\frac{R}{1+L}}$ for all $r \in [0, \sqrt{L}]$ and $R \geq 0$, and $I_0\left(\frac{2\sqrt{\mathsf{K}R}\, r}{1+r^2}\right) \leq I_0(\sqrt{\mathsf{K}R}) \leq e^{\sqrt{\mathsf{K}R}}\ \forall R, r \geq 0$. Finally, (142) follows from the observation that the integrand in (141) is less than 1 $\forall r, \mathsf{K} \geq 0$. Note that the upper bound $u(R, \mathsf{K})$ is not a function of F_n and decreases exponentially in R for sufficiently large values of R. Next, we find the following upper bound:

$$\left| \frac{R^{m-2}}{(m-2)!}\, g(R, F_n, \mathsf{K}) \log\, g(R, F_n, \mathsf{K}) \right| \leq \frac{R^{m-2}}{(m-2)!} \left(4\, g^{0.9}(R, F_n, \mathsf{K}) + g^2(R, F_n, \mathsf{K})\right) \tag{144}$$

$$\leq \frac{R^{m-2}}{(m-2)!} \left(4\, u^{0.9}(R, \mathsf{K}) + u^2(R, \mathsf{K})\right) \quad \forall R, \mathsf{K} \geq 0. \tag{145}$$

(144) follows from the fact that $|x \log(x)| \leq 4x^{0.9} + x^2$ for all $x \geq 0$, and (145) follows from (143). Note that the upper bound in (145) does not depend on F_n and is integrable due to the exponential decay of $u(R, \mathsf{K})$ in R for sufficiently large values of R. Applying the Dominated Convergence Theorem with the upper bound in (145) justifies (138). We further consider

$$\left| f_{\mathsf{K}}(\mathsf{K}) \int_0^{\infty} dR\, \frac{R^{m-2}}{(m-2)!}\, g(R, F_n, \mathsf{K}) \log\, g(R, F_n, \mathsf{K}) \right| \leq f_{\mathsf{K}}(\mathsf{K}) \int_0^{\infty} dR\, \frac{R^{m-2}}{(m-2)!} \left| g(R, F_n, \mathsf{K}) \log\, g(R, F_n, \mathsf{K}) \right| \tag{146}$$

$$\leq f_{\mathsf{K}}(\mathsf{K}) \int_0^{\infty} dR\, \frac{R^{m-2}}{(m-2)!} \left(4\, g^{0.9}(R, F_n, \mathsf{K}) + g^2(R, F_n, \mathsf{K})\right) \tag{147}$$

Note that $f_{\mathsf{K}}(\mathsf{K}) = \frac{1}{E\{\mathsf{K}\}} e^{-\frac{\mathsf{K}}{E\{\mathsf{K}\}}}$ where $E\{\mathsf{K}\} = \frac{\gamma^2 \delta m P}{N_0}$. The integral of the upper bound $u(R, \mathsf{K})$ with respect to R increases exponentially with K. Hence, we need to find a tighter upper bound. We have

$$g(R, F_n, \mathsf{K}) \leq (m-2) \int_0^{\sqrt{L}} \frac{e^{-\frac{R+\mathsf{K}r^2}{1+r^2}}}{(1+r^2)} I_0\left(\frac{2\sqrt{\mathsf{K}R}\, r}{1+r^2}\right) dF_n(r) \tag{148}$$

$$\leq (m-2) \int_0^{\sqrt{L}} e^{-\frac{R+\mathsf{K}r^2 - 2\sqrt{\mathsf{K}R}\, r}{1+r^2}} dF_n(r) \tag{149}$$

$$\leq (m-2) \int_0^{\sqrt{L}} e^{-\frac{(\sqrt{R}-\sqrt{\mathsf{K}}\, r)^2}{1+L}} dF_n(r) \tag{150}$$

$$\leq \begin{cases} (m-2) & R \leq \mathsf{K}L \\ (m-2) e^{-\frac{(\sqrt{R}-\sqrt{\mathsf{K}L})^2}{1+L}} & R > \mathsf{K}L \end{cases} \tag{151}$$

$$\triangleq v(R, \mathsf{K}) \quad \forall n, \quad \forall R, \mathsf{K} \geq 0 \tag{152}$$

where (149) follows from the fact that $I_0(x) \leq e^x$, and (150) follows by choosing the largest value $r = \sqrt{L}$ in the denominator of the exponential function. (151) is obtained by noting that $(\sqrt{R} - \sqrt{\mathsf{K}}\, r)^2$ is a nonnegative quadratic function of r, minimized at $r = \sqrt{\frac{R}{K}}$. Hence, if $L \geq \frac{R}{K}$, the minimum value of the quadratic function is zero. Otherwise, it is $(\sqrt{R} - \sqrt{\mathsf{K}L})^2$. From (147) and (152), we have

$$\left| f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, g(R, F_n, \mathsf{K}) \log\, g(R, F_n, \mathsf{K}) \right| \leq f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty dR\, \frac{R^{m-2}}{(m-2)!}\, (4\, v^{0.9}(R, \mathsf{K}) + v^2(R, \mathsf{K})). \tag{153}$$

Note that the upper bound in (153) is independent of F_n. It can also be verified easily that this upper bound is integrable with respect to K due to the facts that f_{K} decreases exponentially with K while the integral in the upper bound produces a result that is at most polynomial in K. Applying the Dominated Convergence Theorem with the integrable upper bound in (153) justifies (137). Hence, the proof is complete.

2) Sufficient and Necessary Kuhn-Tucker Condition: The proof of the sufficient and necessary condition in (70) follows along the same lines as those in [4] and [16]. The weak derivative of $I(\cdot|\hat{h})$ at F_0 is defined as

$$I'_{F_0}(F|\hat{h}) \triangleq \lim_{\theta \to 0} \frac{I[(1-\theta)F_0 + \theta F|\hat{h}] - I(F_0|\hat{h})}{\theta}. \tag{154}$$

The weak derivative of the mutual information in (66) is obtained as

$$I'_{F_0}(F|\hat{h}) = E_{\mathsf{K}} \left\{ \int dF_0(r) \int_0^\infty f_{R|r,\mathsf{K}}(R|r, \mathsf{K}) \log\, g(R, F_0, \mathsf{K})\, dR \right\}$$
$$- E_{\mathsf{K}} \left\{ \int dF(r) \int_0^\infty f_{R|r,\mathsf{K}}(R|r, \mathsf{K}) \log\, g(R, F_0, \mathsf{K})\, dR \right\} + \int dF_0(r) \log(1+r^2) - \int dF(r) \log(1+r^2). \tag{155}$$

Note that if F_0 is indeed the maximizing distribution and hence capacity achieving, then $I'_{F_0}(F|\hat{h}) \leq 0$ for all F satisfying the peak power constraint. Then using the same steps in [4, Appendix II, Theorem 4], we can show that F_0 is a capacity-achieving input distribution if and only if

$$E_{\mathsf{K}}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|r,\mathsf{K})\log\, g(R,F_0,\mathsf{K})\,dR\right\} + \log(1+r^2) + mC_\delta + m - 1 \geq 0 \quad \forall r \in [0,\sqrt{L}] \tag{156}$$

with equality at the points of increase of distribution F_0.

C. Analyticity of the Kuhn-Tucker Condition in the Complex Domain

We consider the following function which is the left-hand-side of the Kuhn-Tucker condition (70) in the complex domain:

$$\Phi(z) = E_{\mathsf{K}}\left\{\int_0^\infty f_{R|r,\mathsf{K}}(R|z,\mathsf{K})\log\, g(R,F_r,\mathsf{K})\,dR\right\} + \log(1+z) + mC_\delta + (m-1). \tag{157}$$

Note that $\log(1+z)$ is an analytic function of $z = z_r + jz_i$ in the entire complex plane excluding the real axis with $z_r \leq -1$ because the principle branch of the logarithm is not analytic only on the negative real line. Next, we investigate the region in which the first term of (157) is analytic. We first note the Differentiation Lemma.

Differentiation Lemma 1: [42, Sec. XII] Let I be an interval of real numbers, possibly infinite. Let U be an open set of complex numbers. Let $f = f(t,z)$ be a continuous function on $I \times U$. Assume:

(i) For each compact subset K of U the integral $\int_I f(t,z)\,\mathrm{d}t$ is uniformly convergent for $s \in K$.

(ii) For each t the function $z \mapsto f(t,z)$ is analytic. Let $F(z) = \int_I f(t,z)\,\mathrm{d}t$.

Then F is analytic on U and $F'(z) = \int_I Df(t,z)\,\mathrm{d}t$ where D is the differentiation operator. Furthermore $Df(t,z)$ satisfies the same hypothesis as f. □

The integral $\int_0^\infty f(t,z)\,\mathrm{d}t$ is said to be uniformly convergent [42] for $z \in K$ if, given $\epsilon > 0$, there exists B_0 such that if $B_0 < B_1 < B_2$, then $\left|\int_{B_1}^{B_2} f(t,z)\,\mathrm{d}t\right| < \epsilon$. From this definition it can be easily shown that if $\int_0^\infty |f(t,z)|\,dt < \infty$, then $\int_0^\infty f(t,z)\,\mathrm{d}t$ is uniformly convergent.

The function

$$f_{R|r,\mathsf{K}}(R|z,\mathsf{K}) = \frac{R^{m-2}}{(m-3)!}\,\frac{e^{-R-\frac{\mathsf{K}z^2}{1+z^2}}}{1+z^2}\int_0^1 (1-a)^{m-3} e^{\frac{az^2R}{1+z^2}}\, I_0\left(\frac{2\sqrt{\mathsf{K}R}\,z\sqrt{a}}{1+z^2}\right) da \tag{158}$$

is analytic in the entire complex plane excluding the points at $z = \pm j$ because rational functions are analytic everywhere except at the points that make the denominator zero; the exponential function and I_0 are analytic everywhere because they can be expanded as power series; and if g and f are analytic then $g \circ f$ is also analytic in the corresponding region. The analyticity of the integral in (158) can also be easily verified using the

Differentiation Lemma since the integration is over a finite interval. In order to find the region in which the the first term on the right-hand-side of (157) is analytic, we need to find the region $\mathcal{D}$ that satisfies for all $z \in \mathcal{D}$

$$\int_0^\infty f_{\mathsf{K}}(\mathsf{K}) \int_0^\infty |f_{R|r,\mathsf{K}}(R|z,\mathsf{K})||\log\, g(R,F_r,\mathsf{K})|\, dR < \infty. \tag{159}$$

We consider

$$|f_{R|r,\mathsf{K}}(R|z,\mathsf{K})| = \left| \frac{R^{m-2}}{(m-3)!} \frac{e^{-\frac{\mathsf{K}z^2}{1+z^2}}}{1+z^2} \int_0^1 (1-a)^{m-3} e^{-R\frac{1+(1-a)z^2}{1+z^2}}\, I_0\left(\frac{2\sqrt{\mathsf{K}R}\, z\sqrt{a}}{1+z^2}\right) da \right| \tag{160}$$

$$\leq \frac{R^{m-2}}{(m-3)!} \frac{\left|e^{-\frac{\mathsf{K}z^2}{1+z^2}}\right|}{|1+z^2|} \int_0^1 (1-a)^{m-3} \left|e^{-R\frac{1+(1-a)z^2}{1+z^2}}\right| \left| I_0\left(\frac{2\sqrt{\mathsf{K}R}\, z\sqrt{a}}{1+z^2}\right)\right| da \tag{161}$$

$$\leq \frac{R^{m-2}}{(m-3)!} \frac{e^{-\Re\left\{\frac{\mathsf{K}z^2}{1+z^2}\right\}}}{|1+z^2|} \int_0^1 (1-a)^{m-3} e^{-R\,\Re\left\{\frac{1+(1-a)z^2}{1+z^2}\right\}}\, I_0\left(2\sqrt{\mathsf{K}R}\sqrt{a}\,\Re\left\{\frac{z}{1+z^2}\right\}\right) da \tag{162}$$

$$\leq \frac{R^{m-2}}{(m-3)!} \frac{e^{-\Re\left\{\frac{\mathsf{K}z^2}{1+z^2}\right\}}}{|1+z^2|} \int_0^1 (1-a)^{m-3} e^{-R\,\Re\left\{\frac{1+(1-a)z^2}{1+z^2}\right\}}\, e^{2\sqrt{\mathsf{K}R}\sqrt{a}\left|\Re\left\{\frac{z}{1+z^2}\right\}\right|} da \tag{163}$$

$$\leq \frac{R^{m-2}}{(m-3)!} \frac{e^{-\Re\left\{\frac{\mathsf{K}z^2}{1+z^2}\right\}}}{|1+z^2|} \int_0^1 (1-a)^{m-3} e^{-R\,\frac{1+z_r^2-z_i^2}{|1+z^2|^2}}\, e^{2\sqrt{\mathsf{K}R}\left|\Re\left\{\frac{z}{1+z^2}\right\}\right|} da \tag{164}$$

$$= \frac{R^{m-2}}{(m-2)!} \frac{e^{-\Re\left\{\frac{\mathsf{K}z^2}{1+z^2}\right\}}}{|1+z^2|}\, e^{-R\,\frac{1+z_r^2-z_i^2}{|1+z^2|^2}}\, e^{2\sqrt{\mathsf{K}R}\left|\Re\left\{\frac{z}{1+z^2}\right\}\right|} \tag{165}$$

$$= \frac{R^{m-2}}{(m-2)!} \frac{e^{-\mathsf{K}\frac{(z_r^2-z_i^2)(1+z_r^2-z_i^2)+4z_r^2z_i^2}{|1+z^2|^2}}}{|1+z^2|}\, e^{-R\,\frac{1+z_r^2-z_i^2}{|1+z^2|^2}}\, e^{2\sqrt{\mathsf{K}R}\,\frac{|z_r(1+z_r^2-z_i^2)+2z_rz_i^2|}{|1+z^2|^2}} \tag{166}$$

$$= \frac{R^{m-2}}{(m-2)!} \frac{e^{-\mathsf{K}\frac{(z_r^2-z_i^2)(1+z_r^2-z_i^2)+4z_r^2z_i^2}{|1+z^2|^2}}}{|1+z^2|}\, e^{-\frac{\left(\sqrt{R(1+z_r^2-z_i^2)}-\sqrt{\mathsf{K}}\,\frac{|z_r(1+z_r^2-z_i^2)+2z_rz_i^2|}{\sqrt{1+z_r^2-z_i^2}}\right)^2}{|1+z^2|^2}}\, e^{\mathsf{K}\frac{(z_r(1+z_r^2-z_i^2)+2z_rz_i^2)^2}{(1+z_r^2-z_i^2)|1+z^2|^2}} \tag{167}$$

$$= \frac{R^{m-2}}{(m-2)!} \frac{e^{-\mathsf{K}\left(\frac{(z_r^2-z_i^2)(1+z_r^2-z_i^2)+4z_r^2z_i^2}{|1+z^2|^2}-\frac{(z_r(1+z_r^2-z_i^2)+2z_rz_i^2)^2}{(1+z_r^2-z_i^2)|1+z^2|^2}\right)}}{|1+z^2|}\, e^{-\frac{\left(\sqrt{R(1+z_r^2-z_i^2)}-\sqrt{\mathsf{K}}\,\frac{|z_r(1+z_r^2-z_i^2)+2z_rz_i^2|}{\sqrt{1+z_r^2-z_i^2}}\right)^2}{|1+z^2|^2}} \tag{168}$$

$$= \frac{R^{m-2}}{(m-2)!} \frac{e^{\mathsf{K}\left(\frac{z_i^2(1+z_r^2-z_i^2)+\frac{4z_r^2z_i^4}{1+z_r^2-z_i^2}}{|1+z^2|^2}\right)}}{|1+z^2|}\, e^{-\frac{\left(\sqrt{R(1+z_r^2-z_i^2)}-\sqrt{\mathsf{K}}\,\frac{|z_r(1+z_r^2-z_i^2)+2z_rz_i^2|}{\sqrt{1+z_r^2-z_i^2}}\right)^2}{|1+z^2|^2}} \tag{169}$$

In the above formulations, $\Re(z)$ denotes the real value of the complex-valued number $z = z_r + jz_i$ whose real and imaginary components are also denoted by z_r and z_i, respectively. (161) follows by taking the absolute value of the integrand instead of the absolute value of the integral. (162) follows from the facts that $\Re\left(e^z\right) = e^{\Re(z)}$ and $|I_0(z)| \leq I_0(\Re(z))$. (163) is due to $I_0(x) \leq e^{|x|}$ for a real number x. (164) is obtained from the

bounds $\Re\left\{\frac{1+(1-a)z^2}{1+z^2}\right\} \geq \frac{1+z_r^2-z_i^2}{|1+z^2|^2}$ which holds for all $a \in [0,1]$ and $|z_r| \geq |z_i|$, and $e^{2\sqrt{\mathsf{K}R}\sqrt{a}\left|\Re\left\{\frac{z}{1+z^2}\right\}\right|} \leq e^{2\sqrt{\mathsf{K}R}\left|\Re\left\{\frac{z}{1+z^2}\right\}\right|}$ $\forall a \in [0,1]$. (165) follows by evaluating the integral in (164), in which the only term that depends on a is $(1-a)^{m-3}$. (166) is obtained by explicitly expressing $\Re\left\{\frac{\mathsf{K}z^2}{1+z^2}\right\}$ and $\Re\left\{\frac{z}{1+z^2}\right\}$ in terms of z_r and z_i, the real and imaginary components of z. (167) follows by expressing the exponents of the second and third exponential functions as a quadratic function of $\sqrt{R}$. Eventually, (169) is obtained from straightforward algebraic computations.

The following lower bound on $g(R, F_r, \mathsf{K})$ can easily be verified by noting that $e^x \geq 1$ and $I_0(x) \geq 1$ for all $x \geq 0$:

$$g(R, F_r, \mathsf{K}) \geq e^{-R} \int_0^L \frac{e^{-\frac{\mathsf{K}r^2}{1+r^2}}}{1+r^2}\, dF_r \geq e^{-R}\, \frac{e^{-\frac{\mathsf{K}L}{1+L}}}{1+L}. \tag{170}$$

From the above lower bound, we see that $|\log g(R, F_r, \mathsf{K})|$ increases at most linearly in both R and K for sufficiently large values of R and K. Therefore, if $(1 + z_r^2 - z_i^2) > 0$, then the upper bound in (169) decreases exponentially in R, and as a result, the inner integral in (159) converges. This condition is satisfied in the region where $|z_r| \geq |z_i|$.

Note that the upper bound in (169) increases exponentially in K. However, the value of the function

$$c(z_r, z_i) = \frac{z_i^2(1 + z_r^2 - z_i^2) + \frac{4z_r^2 z_i^4}{1+z_r^2-z_i^2}}{|1+z^2|^2} \tag{171}$$

can be made arbitrarily small by choosing arbitrarily small values for $|z_i|$. Note also that $f_{\mathsf{K}}(\mathsf{K}) = \frac{1}{E\{\mathsf{K}\}} e^{-\frac{\mathsf{K}}{E\{\mathsf{K}\}}}$ where $E\{\mathsf{K}\} = \frac{\gamma^2 \delta m P}{N_0}$. Hence, in the region where $c(z_r, z_i) < \frac{N_0}{\gamma^2 \delta m P}$, we have the integrand in (159) exponentially decreasing in K and as a result the integral converges. it can be shown that for a fixed $|z_i| < 1$, $c(z_r, z_i)$ is a monotonically decreasing function of $z_r \geq 0$ achieving its maximum of at $\frac{z_i^2}{1-z_i^2}$ at $z_r = 0$. Hence, we consider the following region in the complex domain:

$$\begin{aligned}\mathcal{D} = &\left\{(z_r, z_i) : 0 \leq z_r \leq \min\left(\frac{1}{\sqrt{2}}, \sqrt{\frac{N_0}{2\gamma^2 \delta m P}}\right) \text{ and } |z_i| \leq z_r\right\} \\ &\bigcup \left\{(z_r, z_i) : z_r > \min\left(\frac{1}{\sqrt{2}}, \sqrt{\frac{N_0}{2\gamma^2 \delta m P}}\right) \text{ and } |z_i| \leq \min\left(\frac{1}{\sqrt{2}}, \sqrt{\frac{N_0}{2\gamma^2 \delta m P}}\right)\right\}\end{aligned} \tag{172}$$

In region $\mathcal{D}$, $c(z_r, z_i) < \frac{N_0}{\gamma^2 \delta m P}$ and $|z_r| \geq |z_i|$. Hence, the integral in (159) converges in this region. Moreover, this region includes the positive real line.

REFERENCES

[1] T. Ericsson, "A Gaussian channel with slow fading," *IEEE Trans. Inform. Theory*, vol. 16, pp. 353-355, May 1970.

[2] L. H. Ozarow, S. Shamai, and A. D. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Vehicular Technology*, vol. 43, pp. 359-377, May 1994.

[3] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1986-1992, Nov. 1997.

[4] I. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh fading channels," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1290-1301, May 2001.

[5] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 139-157, Jan. 1999.

[6] T. L. Marzetta and B. M. Hochwald, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, pp. 543-564, Mar. 2000.

[7] J. Huang and S. P. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Trans. Inform. Theory*, vol. 51, pp. 2336-2351, July 2005.

[8] T. H. Chan, S. Hranilovic and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inform. Theory*, vol. 51, pp. 2073-2088, June 2005.

[9] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2426-2467, Oct. 2003.

[10] L. Zheng and D. N. C. Tse, "Communication on the Grassman manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Trans. Inform. Theory*, vol. 48, pp. 359-383, Feb. 2002.

[11] L. Zheng, D. N. C. Tse, and M. Médard "Channel coherence in the low SNR regime," *IEEE Trans. Inform. Theory*, vol. 53, pp. 976-997, March 2007.

[12] A. Lapidoth and S. Shamai (Shitz), "Fading channels: How perfect need 'perfect side information' be?," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1118-1134, May 2002.

[13] M. Médard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of channel," *IEEE Trans. Inform. Theory*, vol. 46, pp. 933-946, May 2000.

[14] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1319-1343, June 2002.

[15] M. Médard and R. G. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Trans. Inform. Theory*, vol. 48, pp. 840-852, Apr. 2002.

[16] M. C. Gursoy, H. V. Poor, and S. Verdú, "The noncoherent Rician fading channel – Part I : Structure of the capacity-achieving input," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2193-2206, Sept. 2005.

[17] M. C. Gursoy, H. V. Poor, and S. Verdú, "The noncoherent Rician fading channel – Part II : Spectral efficiency in the low power regime," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2207-2221, Sept. 2005.

[18] M. C. Gursoy, H. V. Poor and S. Verdú, "Spectral Efficiency of Peak Power Limited Rician Block-Fading Channels," *Proc. 2004 IEEE Int'l. Symp. Inform. Theory*, Chicago, IL, June 27 - July 2, 2004.

[19] V. G. Subramanian and B. Hajek, "Broad-band fading channels: signal burstiness and capacity," *IEEE Trans. Inform. Theory*, vol. 48, pp. 809-827, Apr. 2002.

[20] I. E. Telatar and D. N. C. Tse, "Capacity and mutual information of wideband multipath fading channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1384-1400, July 2000.

[21] V. Sethuraman and B. Hajek, "Capacity Per Unit Energy of Fading Channels With a Peak Constraint," *IEEE Trans. Inform. Theory*, vol. 51, pp. 3102-3120, Sept. 2005.

[22] L. Tong, B. M. Sadler, and M. Dong, "Pilot-assisted wireless transmission," *IEEE Signal Processing Mag.*, pp. 12-25, Nov. 2004.

[23] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links," *IEEE Trans. Inform. Theory*, vol. 49, pp. 951-963, Apr. 2003.

[24] J. Baltersee, G. Fock, and H. Meyr, "An information theoretic foundation of synchronized detection," *IEEE Trans. Commun.*, vol. 49, pp. 2115-2123, Dec. 2001.

[25] J. Baltersee, G. Fock, and H. Meyr, "Achievable rate of MIMO Channels with data-aided channel estimation and perfect interleaving," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2358-2368, Dec. 2001.

[26] D. Samardzija and N. Mandayam, "Pilot-assisted estimation of MIMO fading channel response and achievable data rates," *IEEE Trans. Signal Process.*, vol. 51, pp. 2882-2890, Nov. 2003.

[27] S. Ohno and G. B. Giannakis, "Average-rate optimal PSAM transmissions over time-selective fading channels," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 712-720, Oct. 2002.

[28] S. Misra, A. Swami, and L. Tong, "Optimal training for time-selective wireless fading channels using cutoff rate," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1-15, 2006.

[29] X. Deng and A. M. Haimovich, "Achievable rates over time-varying Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 55, pp. 1397-1406, July 2007.

[30] I. Abou-Faycal, M. Médard, and U. Madhow, "Binary adaptive coded pilot symbol assisted modulation over Rayleigh fading channels without feedback," *IEEE Trans. Commun.*, vol. 53, pp. 1036-1046, June 2005.

[31] S. Furrer and D. Dahlhaus, "Multiple-antenna signaling over fading channels with estimated channel state information: Capacity Analysis," *IEEE Trans. Inform. Theory*, vol. 53, pp. 2028-2043, June 2007.

[32] S. Adireddy, L. Tong, and H. Viswanathan, "Optimal placement of training for frequency-selective block-fading channels," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2338-2353, Aug. 2002.

[33] M. F. Sencan and M.C. Gursoy, "Achievable rates for pilot-assisted transmission over Rayleigh fading channels," Proceedings of the 40th Annual Conference on Information Sciences and Systems, Prnceton University, Princeton, NJ, March, 22-24, 2006.

[34] S. Akin and M. C. Gursoy, "Training optimization for Gauss-Markov Rayleigh fading channels," Proc. of the IEEE International Conference on Communications (ICC), Glasgow, 2007.

[35] S. Akin and M. C. Gursoy, "Pilot-symbol-assisted communications with noncausal and causal Wiener filters," submitted to the IEEE International Conference on Communications (ICC), Beijing, 2008.

[36] J. A. Gansman, M. P. Fitz, and J. V. Krogmeier, "Optimum and suboptimum frame synchronization for pilot-symbol-assisted modulation," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1327-1337, Oct. 1997.

[37] W.-Y. Kuo and M. P. Fitz, "Frequency offset compensation of pilot symbol assisted modulation in frequency flat fading," *IEEE Trans. Commun.*, vol. 45, pp. 1412-1416, Nov. 1997.

[38] G. K. Kaleh, "Channel equalization for block transmission systems," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1728-1736, Jan. 1995.

[39] D. G. Luenberger, *Optimization by Vector Space Methods.* Wiley: New York, 1969

[40] W. Rudin, *Principles of Mathematical Analysis.* McGraw Hill: New York, 1964.

[41] K. Knopp, *Theory of Functions.* Dover: New York, 1945.

[42] S. Lang, *Complex Analysis 2nd. Ed.* Springer-Verlag: New York, 1985.

[43] J. G. Proakis, *Digital Communications 3rd. Ed.* McGraw-Hill: New York, 1995.